Is it Worth to Become an Airbnb Superhost?

Liuyi Pan - 1006211573

December 17, 2021

Abstract

Since 2008, Airbnb has grown rapidly, with exponential growth in the number of listings and competition in the host market. The concept of superhost has also raised more and more concerns, and it has become a question whether it can bring higher revenue. In this paper, a valid sample of 7370 listings in Venice, Italy was downloaded from the Inside Airbnb website. two sample hypothesis test was employed to investigate whether being a superhost has an effect on house prices. Further, using propensity scores matching to see if this effect led to an increase in Airbnb prices. Results showed that Airbnb prices in Venice are indeed influenced by the host status, and those B&Bs with the "Superhost" badge are more likely to have higher prices. The results of this study could help hosts make better decisions and allow both hosts and tenants to benefit from this economic model.

Keywords

Airbnb, Superhost, Propensity Scores Matching, Two Sample Hypothesis Test, Regression Model, Observational Data.

Introduction

Airbnb takes a unique approach to accommodation, offering travelers a home-like B&B rather than a traditional hotel. Since its inception in 2008, it has grown rapidly but has also created a competitive market environment where hosts need to use appropriate pricing to make revenue. In this host market, "Superhost" has become a very hot concept. According to the Airbnb website, to become a superhost, you must meet the following requirements: (1) Completed at least 10 trips or 3 reservations that total at least 100 nights, (2) Maintained a 90% response rate or higher, (3) Maintained a 1% cancellation rate or lower, and (4) Maintained a 4.8 overall rating (Airbnb, 2016).

Thus, it takes a lot of time and effort to become a superhost. For instance, they need to keep improving their facilities and service quality in order to satisfy their guests and get high ratings. But whether this title leads to higher revenue and whether it is worthwhile for owners to make the above efforts to earn and retain it has been a growing number of hosts' doubts. Therefore, the research question in this paper is **whether superhosts have an impact on house prices and whether this impact increases house prices?** In addition to having referential value for host making decisions, this study also helps renters understand Airbnb price bias, which in turn helps them make better decisions when comparing and choosing accommodations on Airbnb.

Also, my hypothesis here is that Airbnb prices are indeed influenced by the superhost status, and that accommodations owned by the superhost are more likely to have higher prices.

Data

Data Collection

The data for this report was published on Inside Airbnb website (Cox, 2018), which has collected publicly available information from the Airbnb website and has done some analysis, cleaning and aggregation to facilitate public discussion. Listing data on the site is categorized by major cities around the world. For this study, a dataset of 7370 Airbnb listings in Venice, Italy, as of November 4, 2021 was downloaded. Therefore, listings added after this date will not be captured by our analysis. Here we chose Venice as our study setting because it is one of the top tourist destinations in the world. Research shows that, in just seven years, the number of B&Bs and rooms to rent in Venice has risen 1008 percent (Da Mosto et al., 2009).

Data Clean

Looking specifically at the data, it contains many basic features of listings. But not all the variables in the dataset are needed in our study. First, we selected the two variables of most interest:

- price(euros): the daily price of the Airbnb,
- host is superhost: whether the Airbnb's host is a superhost ,"t" means yes and "f" means no.

Then the three factors mentioned in the introduction that influence host to be superhost were selected for the subsequent construction of the model:

- host response rate: The response rate of the Airbnb's host;
- review scores rating: The review scores rating of the Airbnb, ranging from 0 to 5.

The research by Chen and Xie on Airbnb price prediction shows that the functional characteristics of Airbnb listings are significantly associated to the price of the listings (Chen & Xie, 2017). Thus, we also chose the following variables in the dataset to indicate the intrinsic features of the listings:

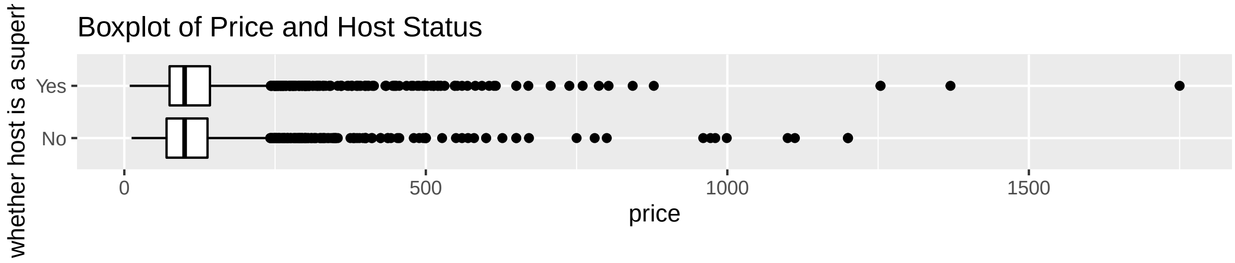
- room type: The room type of the Airbnb, including "Entire home/apt", "Hotel room", "Private room", and "Shared room";
- accommodates: The maximum capacity of the Airbnb;
- bathrooms text: The number of bathrooms in the Airbnb using textual description, i.e. "2 baths";
- bedrooms: The number of bedrooms in the Airbnb.

Note that both "host response rate" and "host acceptance rate" are expressed in text with percentage signs, so we removed the percentage signs and converted them to numeric form. Additionally, the "bathrooms text" variable is also recorded in text form, so we extracted its number and stored it in the new "bathrooms" variable. Finally, removing all missing values and the original "bathrooms text" variable, we obtained 4992 observations with 9 variables describing the basic information of the listing. The first 10 rows can be checked in Appendix Table 5.

Data Summary

Now we perform further analysis of these variables. Firstly, we checked whether superhost would set different house prices or not. Through the boxplot in Figure 1 we cannot observe a huge difference between the prices set by superhost and non-superhost, but we can find from the summary table in Figure 1 that the average price set by superhost (€128.307) is higher than that of non-superhost (€118.481), and with a median value of €100 for both. Moreover, the highest price set by the superhost has 1750 Euros, while the one set by the non-superhost is 550 Euros less.

Next, we check the average price for each room type, and we see from Figure 2 that except for the entire home type, the non-superhost set higher prices for the hotel room and the private room, which is not the same as our hypothesis. For the shared room, we cannot draw any conclusion because there is no example of superhost‘s listing.



Summary Table of Airbnb Price by Host Status

host_is_superhost	Number	Median	Mean	Min	Max
No	2657	100	118.481	12	1200
Yes	2335	100	128.307	9	1750

Figure 1: Host Status and Price Analysis

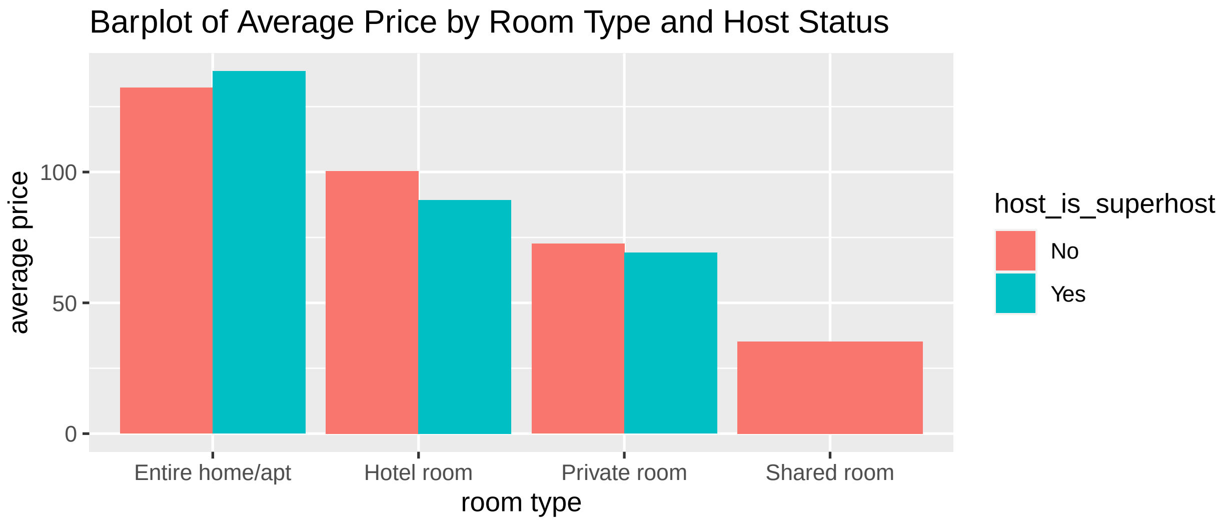


Figure 2: Room type and Price Analysis by Host Status

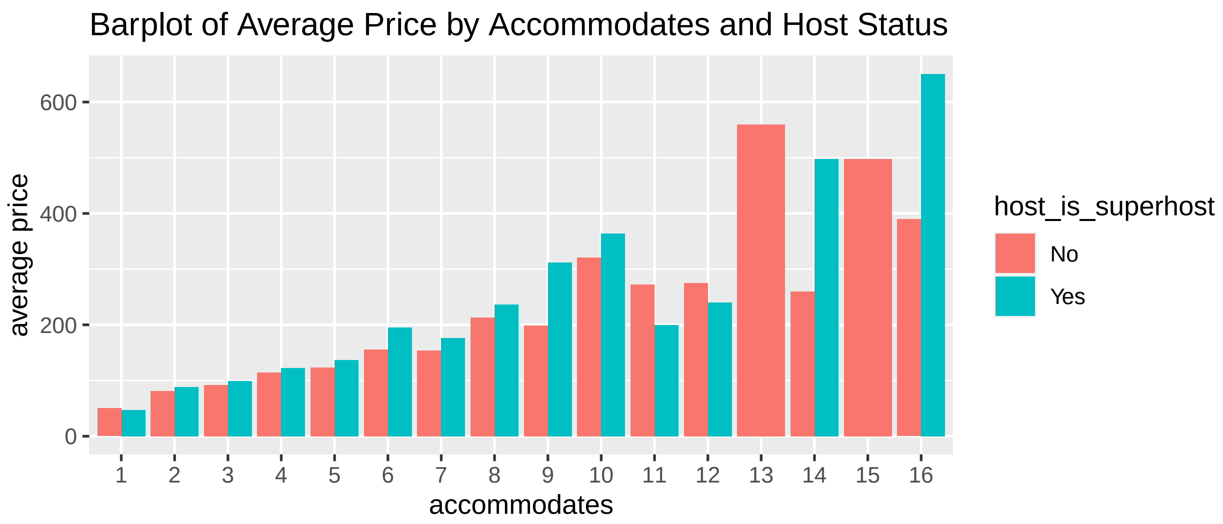


Figure 3: Accommodates and Price Analysis by Host Status

Finally, consider the maximum capacity of Airbnb. As shown in Figure 3, most superhosts set higher prices for listings regardless of capacity size, and only for listings with accommodates of 11,12 people, non-superhosts set higher prices.

The above analysis shows that superhosts' listings seem to have higher price, but there are still situations that violate our hypothesis, so next we will use hypothesis test and propensity scores matching method to explore the research question further.

Methods

Two Sample T-Test

The first statistical method used is two sample t-test, which can test our first research question: whether the unknown population means of Airbnb price between two groups(superhost vs. non-superhost) are equal or not. First, we identify a null hypothesis and an alternative hypothesis, which is the hypothesis we believe and an alternative to the null hypothesis respectively. Then in the hypothesis testing process, we try to find the probability of what we observe if the null hypothesis is applyed. The more unlikely we observe, the stronger the evidence against the null hypothesis.

Based on our research question, we want to know we wanted to know if the average price of the superhost listing differs from that of the non-superhost. So we perform a hypothesis test with the following hypothesis:

$$H_0 : \mu_{superhost} = \mu_{non.superhost} \quad vs. \quad H_a : \mu_{superhost} \neq \mu_{non.superhost}$$

where $\mu_{superhost}$ reflects the population mean of Airbnb Price for superhosts, $\mu_{non.superhost}$ reflects the population mean of Airbnb Price for non_superhosts.

The test statistic of this hypothesis test is calculated by:

$$T = \frac{\bar{X_1} - \bar{X_2}}{\sqrt{s^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

where $\bar{X_1}$ is the sample mean price of the superhost group, $\bar{X_1}$ is the sample mean price of the non-superhost group, n_1 is the sample size of the superhost group, n_2 is the sample size of the non-superhost group and $s^2 = \frac{\sum_{i=1}^{n_1}(X_i - \bar{X_1})^2 + \sum_{i=1}^{n_2}(X_i - \bar{X_2})^2)}{n_1+n_2-2}$,

The corresponding P-value can be subsequently derived as the probability that the difference between the two groups of sample mean is more extreme than the test statistic when H_0 is true, which can be computed as $P(|t_{n_1+n_2-2}| > |T|)$. Statisticians usually use a 5% significance level as a threshold: if the p-value is less than 5% then we have enough evidence to reject the null hypothesis, and vice versa we fail to reject the null hypothesis.

The two sample t-test is appropriate firstly because house price is a continuous variable, which means that we are able to establish a null hypothesis. Futhermore, since our sample is randomly selected and has a big sample size, we can say the two sample is independent and the data follow the normal probability distribution, which also fits the assumptions.

Propensity Scores Matching

The second statistical method used is propensity scores matching(PSM), which can check our second research question: if being a superhost or not has a causal effect on their Airbnb price. The basic idea of PSM is to reduce the influence of confounding variables on response, so that we can better assess whether the treatment really "influences" the response. In this study, the response is Airbnb price and the treatment is being a superhost. More specifically, we match Airbnb listings with different prices, but similar propensity scores of being a superhost.

Logistic Regression Model

First of all, a logistic model is used to model the proportion of being a superhost since the predicted outcome is a binary response variable. Also, the observations of sample could be independent to meet the assumption. So, we think using logistic regression model is appropriate.

The predictors we apply to the model are three factors(host response rate, acceptance rate and listing review scores) mentioned in the introduction and data section that influence host to be superhost, and four variables(room type, accommodates and the bathrooms and bedrooms number) mentioned in the data section as well that indicate the intrinsic features of the listings. The specific mathematical model is as follows:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{ResponseRate} + \beta_2 x_{AcceptanceRate} + \beta_3 x_{ReviewScores}$$

$$+\beta_4 x_{RoomType} + \beta_5 x_{Accommodates} + \beta_6 x_{Bedrooms} + \beta_7 x_{Bathrooms}$$

where:

- p represents is the probability of being a Airbnb superhost.
- β_0 represents the intercept of the model.
- β_1 represents the slope of response rate. So, for everyone one unit increase in response rate, we expect a β_1 increase log odds of being a superhost.
- β_2 represents the slope of acceptance rate. So, for everyone one unit increase in acceptance rate, we expect a β_2 increase log odds of being a superhost.
- β_3 represents the slope of review scores. So, for everyone one unit increase in review scores, we expect a β_3 increase log odds of being a superhost.
- β_4 represents the average difference in log odds of being a superhost between different room type for a certain response rate, acceptance rate, review scores, accommodates and bathrooms and bedrooms number.
- β_5 represents the slope of accommodates. So, for everyone one unit increase in accommodates, we expect a β_5 increase log odds of being a superhost.
- β_6 represents the slope of the number of bedrooms. So, for everyone one unit increase in bedrooms number, we expect a β_6 increase log odds of being a superhost.
- β_7 represents the slope of the number of bathrooms. So, for everyone one unit increase in bathrooms number, we expect a β_7 increase log odds of being a superhost.

Matching Process

Then we use the prediction in above logstic model to create matched. For every Airbnb listing that are treated (host is superhost), we want the untreated listing that are considered as similar as possible, which means they have similar propensity scores. Here, we use a matching function that can find which is the closet of the ones that are not treated, to each one that is treated. Next, matched treatment is selected and we get a new dataset.

The goal of the matching is to have balance in the treatment group and we need to make sure the observation matched should have similar traits. Hence, we apply two sample hypothesis test again to see if we have a good quality of matching and the traits we picked is the proportion of hotel room. The null hypothesis here is $H_0 : p_{superhost} = p_{non.superhost}$, which means the population proportion of hotel room for superhosts is the same as that of non-superhosts.The rest of the analysis is the similar to above: we will also use p-value to check if we can reject the H_0 or not.

Mulitiple Linear Regression Model

Finally, mulitiple linear regression model are employed to detect the relationship between Airbnb price and being a superhost. The response of this model is the price of Venice's Airbnb listings, the predictor is host status. The specific mathematical model is as follows:

$$Y = \beta_0 + \beta_1 X_{treatment} + \epsilon$$

where:

- β_0 represents the mean value of Airbnb price when the host is a superhost.
- β_1 represents the slope of being a superhost. So the mean value of Airbnb price for the superhost will be β_1 higher than the non-superhost.
- ϵ is the random error term, which explains the difference between the theoretical value of the model and the actual observed results.

After model building, we run F-test to see whether a significant linear relationship exists overall, here our null hypothesis is none of the predictors are linearly related to the response, which is just the same as host status is not linearly related to Airbnb price since we only have one predictor in the model. And we will reject null hypothesis if a p-value is below 0.05, otherwise, we fail to reject that.

Results

Two Sample T-Test

The results of the two sample t-test for checking the difference in population means of Airbnb price between two groups are shown in Table 1. The test statistic foris -3.432105 and corresponding p-value is 0.0006041.

Table 1: Summary of T-Test

statistic	p.value
-3.432105	0.0006041

Propensity Scores Matching

Seven variables are used to conduct a logistic regression model for forecasting the proportin of being a superhost. From the Table 2 below we can see that the p-value for host's response rate, acceptance rate, listing's review scores, accommodates and bathrooms number is really small(less than 0.05).

Table 2: Summary of Logistic Model Coeffient

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-29.5187513	1.2226528	-24.1432008	0.0000000
host_response_rate	0.0803963	0.0086863	9.2555108	0.0000000
host_acceptance_rate	0.0128037	0.0020131	6.3602339	0.0000000
review_scores_rating	4.3364811	0.1783369	24.3162322	0.0000000
room_typeHotel room	0.0278517	0.2161622	0.1288462	0.8974794
room_typePrivate room	-0.5820837	0.0981450	-5.9308546	0.0000000
room_typeShared room	-12.0162946	241.0462976	-0.0498506	0.9602415
accommodates	-0.1435480	0.0282941	-5.0734349	0.0000004
bedrooms	0.0823200	0.0647487	1.2713769	0.2035946
bathrooms	0.1683766	0.0702946	2.3953015	0.0166067

After doing propensity scores matching, we had 2335 treatment and 4670 observations in the new matched dataset. The first 10 rows can be checked in Appendix Table 6.

The results of the two sample z-test for examining if the observation matched have similar traits are shown in Table 3. The test statistic for is 0.6817445 and corresponding p-value is 0.4089867.

Table 3: Summary of Z-Test

statistic	p.value
0.6817445	0.4089867

As can be seen from the Table 4 (below), the final linear regression model that we conduct after propensity scores matching has a very small p-value, 0.0009429.

Table 4: Summary of Mulitiple linear Regression Model

term	estimate	std.error	statistic	p.value
(Intercept)	118.730621	2.046235	58.023956	0.0000000
treatment	9.576017	2.893813	3.309135	0.0009429

All analysis for this report was programmed using `R version 4.0.4`.Packages used for this report include `dplyr`(Hadley et al., 2021), `ggpubr`(Alboukadel, 2020), `ggplot2`(Wickham, 2016), `knitr::kable`(Yihui, 2021), `kableExtra`(Hao, 2021), `broom`(David, Alex, & Simon, 2021), `arm`(Andrew & Yu-Sung, 2021) and `AER`(Christian & Achim, 2008) packages.

Conclusions

From the results of two sample t-test, small p-value indicates that we have evidence to show the average price of the superhost listing differs from that of the non-superhost. This also answers our first research question.

Although not all predictors in logistic regression model is significant to the price, we do want to include these variables because they are highly correlated with the price so that it can be sufficient to balance. And the next two sample z-test, with p-value that is larger than 0.05, suggests we have no evidence to conclude that the population proportion of hotel room for superhosts is the same as that of non-superhosts, thereby our matching is relatively good.

The small p-value of final linear regression model shows that, host status have a significant linear relationship with Airbnb price and the average Airbnb price for superhosts will be €9.576 higher than for non-superhosts. This also answers our second research question.

Taken together, these results suggest that there is an association between Airbnb price and host status, and becoming a super host will, to some extent, lead to an increase in Airbnb price. The results of this study will help hosts make better decisions about becoming a superhost: Hosts should take into account a variety of factors to consider whether the €9 rate increase is worthwhile. At the same time, consumers can benefit from understanding this price difference to help them make better decisions when comparing and choosing accommodations on Airbnb.

Weaknesses and Next Steps

- This paper only studys the Airbnb market in Venice, and the results obtained may vary from region to region because of the different characteristics of B&Bs. Future research can explore the impact of becoming a superhost on the price based on other countries and regions.
- We did not filter out all the features of Airbnb, and there may still be confounding variables that were not "eliminated" effects, which may lead to bias in the final results. Later, we can use more variables to build a logistic regression model to achieve a better balance.
- In checking the quality of matching, we only compared the proportion of hotel rooms between the two different groups. We can subsequently check more variables to make the study more accurate and complete.

Bibliography

1. Airbnb. (2016). *How to Become a Superhost.* Retrieved from https://www.airbnb.ca/help/article/829 /how-to-become-a-superhost

2. Alboukadel Kassambara (2020). *ggpubr: 'ggplot2' Based Publication Ready Plots.* R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr

3. Andrew Gelman and Yu-Sung Su (2021). *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models.* R package version 1.12-2. https://CRAN.R-project.org/package=arm

4. Chen, Y., & Xie, K. (2017). *Consumer valuation of Airbnb listings: A hedonic pricing approach.* International journal of contemporary hospitality management.

5. Christian Kleiber and Achim Zeileis (2008). *Applied Econometrics with R.* New York: Springer-Verlag. ISBN 978-0-387-77316-2. URL https://CRAN.R-project.org/package=AER

6. Cox, M. (2018). *Inside Airbnb.* Retrieved December 15, 2021, from http://insideairbnb.com/get-the-data.html

7. Da Mosto, J., Morel, T., & Gibin, R. (2009). *The Venice Report.* Demography, Tourism, Financing and Change of Use of Buildings.

8. David Robinson, Alex Hayes and Simon Couch (2021). *broom: Convert Statistical Objects into Tidy Tibbles.* R package version 0.7.9. https://CRAN.R-project.org/package=broom

9. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2021). *dplyr: A Grammar of Data Manipulation.* https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.

10. Hao Zhu (2021). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax.* http://haozhu233.github.io/kableExtra/, https://github.com/haozhu233/kableExtra.

11. H. Wickham.(2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York.

12. Yihui Xie (2021). *knitr: A General-Purpose Package for Dynamic Report Generation in R.* R package version 1.34.

Appendix

Ethics Statement

- The website where the data is downloaded, Inside Airbnb, is not associated with or endorsed by Airbnb or any of Airbnb's competitors. And there is no "private" information in the open data. Names, photographs, listings and review details are all publicly displayed on the Airbnb site.
- This paper's research on superhost is not meant to force hosts to all become superhosts, but only to give them and consumers some reference.
- I will keep in mind that my data are not just numbers, but represent real people and situations, and that my work may lead to unintended social consequences due to algorithmic bias.

Supplementary Plots

Table 5: First Ten Rows of Cleaned Data

price	host_is_superhost	host_response_rate	host_acceptance_rate	review_scores_rating	room_type	accommodates	bedrooms	bathrooms
193	f	100	30	4.82	Private room	3	1	1
280	t	100	91	4.83	Entire home/apt	6	3	2
136	t	100	98	4.80	Entire home/apt	5	3	2
150	f	100	0	4.78	Entire home/apt	6	2	1
89	f	100	100	4.73	Entire home/apt	4	2	1
42	f	100	93	4.33	Hotel room	2	1	1
96	t	100	100	4.89	Private room	2	1	1
68	t	100	99	4.81	Private room	2	1	1
130	f	0	0	4.74	Entire home/apt	4	1	2
85	f	100	100	4.50	Entire home/apt	4	1	1

Table 6: First Ten Rows of Matched Data

price	host_is_superhost	host_response_rate	host_acceptance_rate	review_scores_rating	room_type	accommodates	bedrooms	bathrooms	treatment	.fitted
489	t	100	100	0	Entire home/apt	6	2	2.0	1	0.0000000
86	f	100	99	0	Entire home/apt	4	1	1.0	0	0.0000000
180	f	100	97	0	Entire home/apt	4	2	2.0	0	0.0000000
98	t	100	100	0	Hotel room	2	1	1.0	1	0.0000000
82	t	100	33	3	Entire home/apt	4	1	1.0	1	0.0002315
90	f	0	0	5	Private room	2	1	1.5	0	0.0002315