Draft White Paper

(Determine Destination with Data)

DSC 680 Week 3

Blandon S. Lee

12/16/2022

Introduction

Since August 2008, Airbnb has become a multi-Billion-dollar company, and people around the world now mee them as an alternative for booking their travel stays. While other lodging companies are still the first choice of places to stay, Airbnb is becoming more and more popular as the year's pass. The main reason is that renting a whole home or even a private room is better than a hotel.

Currently there are no third-party sites for consumers to get deals on Airbnb lodging. Meaning that they would have to go directly to the Airbnb site and check out reviews and other telling details and determine if the location is acceptable for their needs. However, it does not have to be this way. What if there was a way to filter this data and provide recommendations based on things such as price, reviews, and other defining characteristics. This would better help the traveler to choose an Airbnb as opposed to a hotel or motel.

In this project I am wanting to attempt to identify filters through predictive modeling and determine if it could be used to create an application to be used on the Airbnb sites or bargain sites such as Hotels.com, Travelocity, and other similar sites.

Project Updates

Initially I wanted to answer the following questions

- Does the cost/price of a booking correlate to review ratings?
- Does location impact pricing?
- Does property type impact pricing?

However, as I got into the assignment, I altered my objectives a bit. Not that it was a challenge I just decided to go a slightly different route with the assignment.

I decided to try and answer the following questions instead.

- Does the cost/price of a booking correlate to review ratings?
- What determines Success?
- What variables can help predict if a host will be booked?
- Does having instant booking produce more booking?

Preparation

In this project I am wanting to utilize predictive analytics to better understand the relationships of the data. I feel that this will help in identifying various factors that could assist someone booking an Airbnb site in narrowing their search based on certain specified criteria.

The key factors I thought would be best examined for this project are price of the property, location of the property, the host, the number of properties, the number of reviews, and if the properties can be booked via instant booking.

Methods

Data Understanding

For this project I wanted to start it by looking at and examining the variables within the dataset. To do this I used a few different resources that were available to me such a R, Python, and PowerBi.

- **R**

  Conducted summary statistics to see the trends and what the variables are telling me. I then used a linear regression model to determine what they mean.

- **PowerBI**

  I used the summary statistics and the linear model to produce some visuals for the more prominent variables. I felt that doing this in the beginning would allow me to see the data prior to moving forward with modeling. What I noticed is that price would be a possible target for the models. Other areas of interest were review rating, host response, reviews per month, availability, room type, and guests included.

- **Python**

  I used Python for the modeling the first was K-Modes clustering model; I then two linear regression models and, finally, a decision tree prediction model. These models provide some valuable insights into the questions being asked.

Data Preparation

This portion took up the most time I ended up cleaning the data and then transformed certain categorical predictors into numeric. This was done to situate the data for modeling and analysis. The variables were transformed in Python on the data set by the use of the formula, (df['columnName'] = le.fit_transform(df.columnName.values). Then a new dataset was created and then used in R for the summary statics and linear models.

Modeling

R's initial linear model was for the price variable as a function of the instant book, availability, minimum nights, number of reviews, reviews per month, and host response. These

variables returned some good insights for significance, and all but one returned a p-value under the 0.05 threshold. The overall P-value of the model yielded < 2.2e-16 or (0.00000000000000022) with an F-statistic of 27.68.

In Python I utilized a clustering algorithm to help better understand the data categorical variables. I decided to use K-Modes to deal with the complex variables found in the dataset. I felt that this model would allow matching clusters based on matching categories between each data point. I feel that this model works well when dealing with a dataset that contains a high amount of categorical mixed with numeric data.

I then ran a decision tree model that I felt would use the dataset with a target in mind and return the predicted outcome based on criteria of each node. The mixture of data types should provide the right prediction for a property based on the potential consumer input of yes and no questions.

Analysis

From what I can see the preliminary results that have been generated are promising. Additional analysis will be necessary if definitive conclusions are to be made. Below are the results of the initial models for this project.

```
lm(formula = price ~ instant_bookable + availability_365 + minimum_nights +
    number_of_reviews + reviews_per_month + host_response_rate,
    data = df2)

Residuals:
   Min    1Q Median      3Q     Max
-266.6  -108.8   -50.3   29.9 7820.7

Coefficients:
                      Estimate       Std. Error         t-value      Pr(>|t|)
(Intercept)        278.148607236    9.197847725        30.241    < 2e-16 ***
instant_bookable    -6.950344577    6.050437632        -1.149    0.25070
availability_365    -0.073083931    0.024048295        -3.039    0.00238 **
minimum_nights      -0.000002051    0.000002538        -0.808    0.41904
number_of_reviews   -0.170234684    0.052290551        -3.256    0.00114 **
reviews_per_month  -12.200102032    2.053149427        -5.942    0.00000000294 ***
host_response_rate  -0.272869678    0.098061170        -2.783    0.00541 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 253.7 on 7568 degrees of freedom
Multiple R-squared:  0.02148,           Adjusted R-squared:  0.0207
F-statistic: 27.68 on 6 and 7568 DF,  p-value: < 2.2e-16
```

The linear regression ran in R only provides some of the information needed to determine if predictions can be made. I created a visual using PowerBi and is in Appendix 1 to help illustrate what it might conclude prior to building the model.

The first model that I implemented was K-Modes which does well with categorical data. It will help in taking the entire dataset and returning predictions for each cluster's most common outcomes. Each time the model returned 5 clusters and 3 of those were houses and 2 of them were apartments. The price range returned most often was $100 to $150, but one cluster did return $250.

Building the next models, I used Python. The initial Linear Regression used price as the target with a 60/40 test train split. The result for this Linear Regression is shown in Figure A. It returned predictions for Neighborhood. However, the accuracy was only .26 to .31% meaning that the model was not accurate, and adjustments may need to happen. There could be several different reasons this happened such as linear relationships or poor model selection.

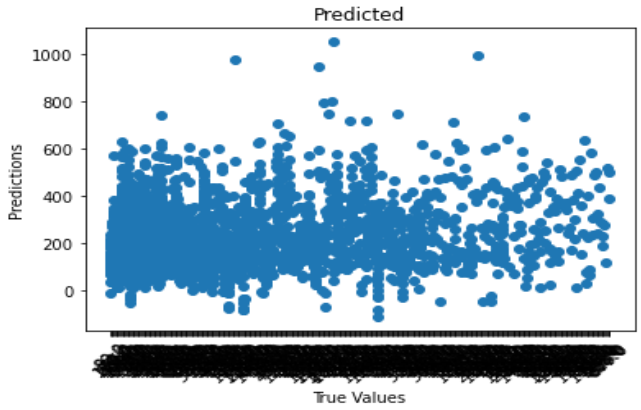


Figure A: Model 1 predictions

The second Linear Regression also used price as the target. However, I believe it used MSE rather than accuracy. The plotted predictions are seen in Figure B returned an MSE of 22110.00. This seems high and will need to be reevaluated.

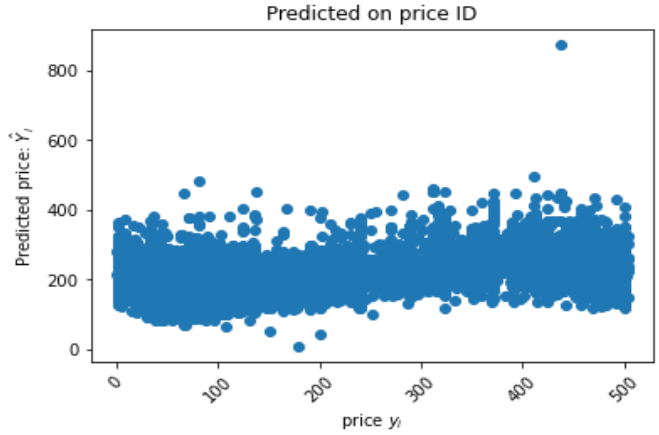


Figure B: Model 2 Predictions

In this model we ran it against two variables one against price and minimum nights. With a 60/40 training test split and a random state of 5. It was able to return an a 25349.17 MSE. Again this is quite high. The residuals for this are seen in Figure C.

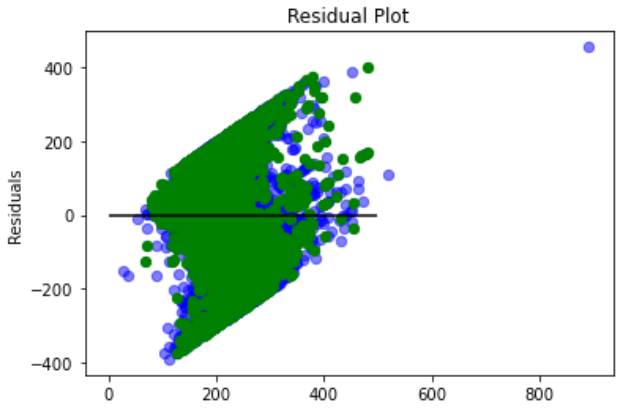


Figure C: Residuals

The final model that was used was the decision tree. This model will provide good predictions for a dataset like the one used in this project. It was set up to identify the neighborhoods that are best suited given the search criteria from the consumer. Looking at Figure D you will see that a host with 8.5 listings and a customer that wants to stay 17 days the search is narrowed down to 371 location prediction.

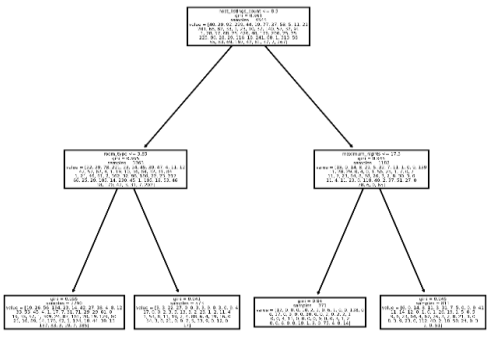


Figure 4: Decision Tree

Conclusion

Using the data for San Francisco Airbnb I was able to garner some useful insights. The analysis performed in this project is just a fraction of the questions I intended to answer. Having more data may provide better results. To better understand the Airbnb industry nationwide or worldwide data would need to be collected and analyzed. What we can see is that the decision tree provided the best prediction for this project. Other models or tweaks to the others may yield better results. However, being able to predict a property based on consumer needs and wants are entirely possible.

References

1. Lilly. (2017). BIGGEST REASONS WHY AIRBNB IS SO POPULAR. The Frugal Gene. Retrieved from https://www.thefrugalgene.com/airbnb-popular/

2. Folger. J. (2020). Airbnb: Advantages and Disadvantages. Investopedia. Retrieved from https://www.investopedia.com/articles/personal-finance/032814/pros-and-cons-using-airbnb.asp

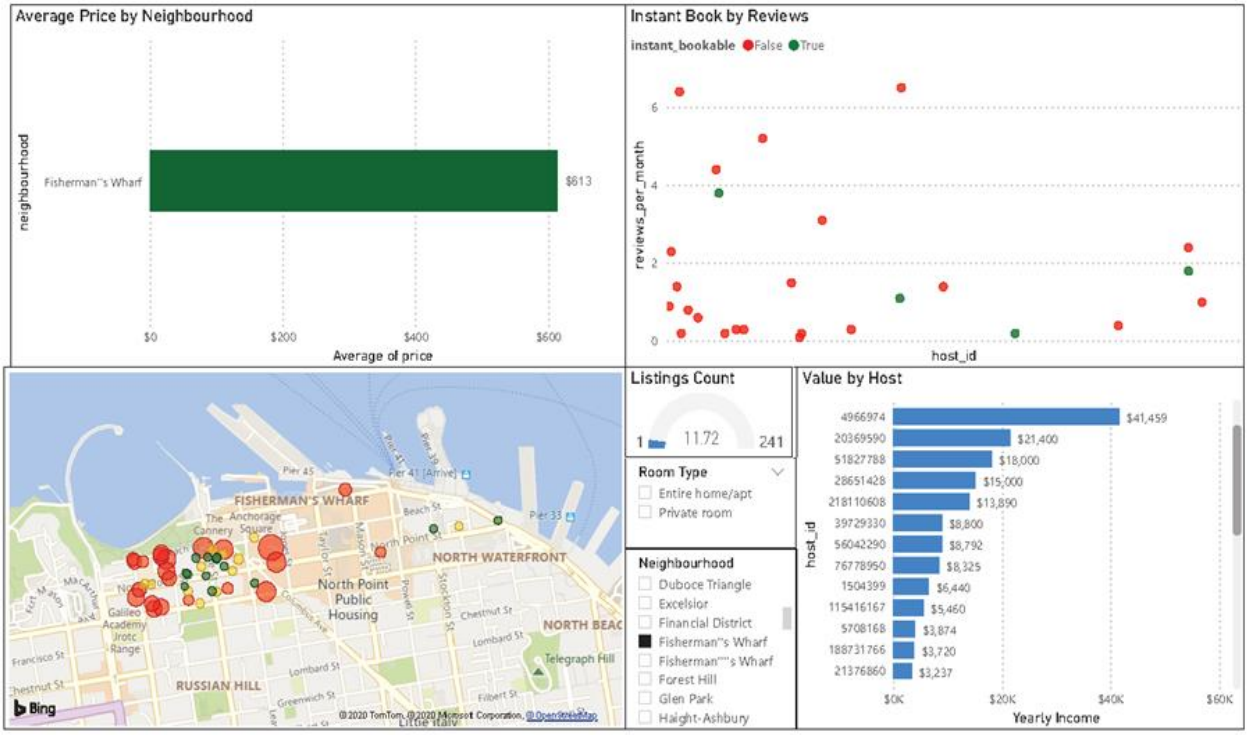
3. Ryerson. (2016). Why tourists choose Airbnb over hotels. Ryerson University. Retrieved from https://www.ryerson.ca/news-events/news/2016/10/why-tourists-choose-airbnb-over-hotels/

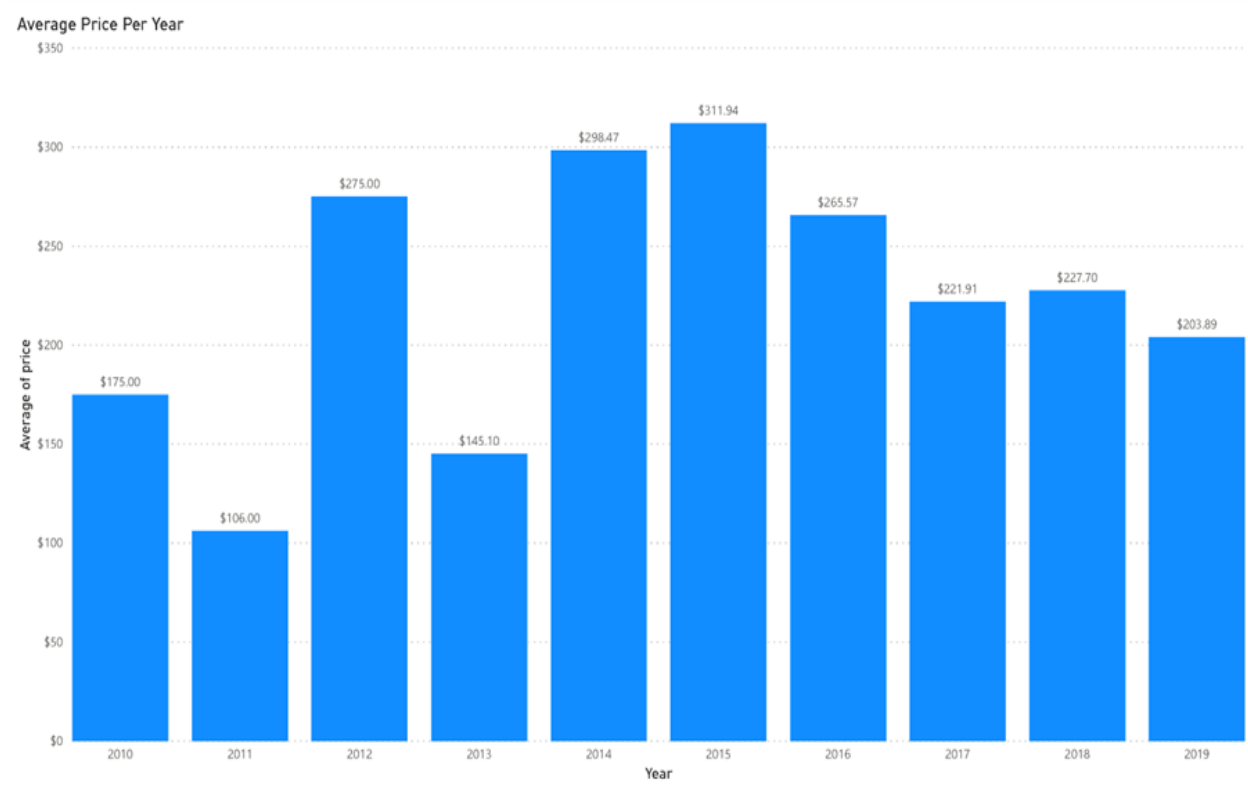
4. Project Pro. (2020). How Data Science increased AirBnB's valuation to $25.5 bn? Retrieved from https://www.dezyre.com/article/how-data-science-increased-airbnbs-valuation-to-25-5-bn/199

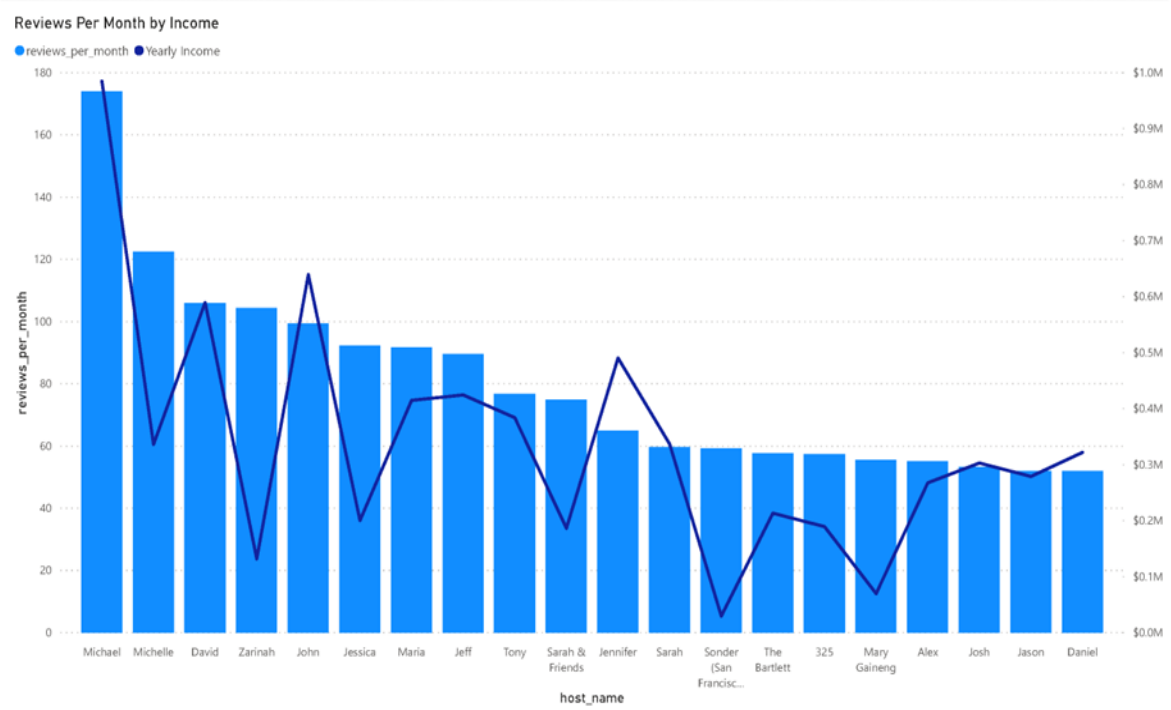
5. Newman, R. (2015). How we scaled data science to all sides of Airbnb over 5 years of hypergrowth. Retrieved from https://venturebeat.com/2015/06/30/how-we-scaled-data-science-to-all-sides-of-airbnb-over-5-years-of-hypergrowth/

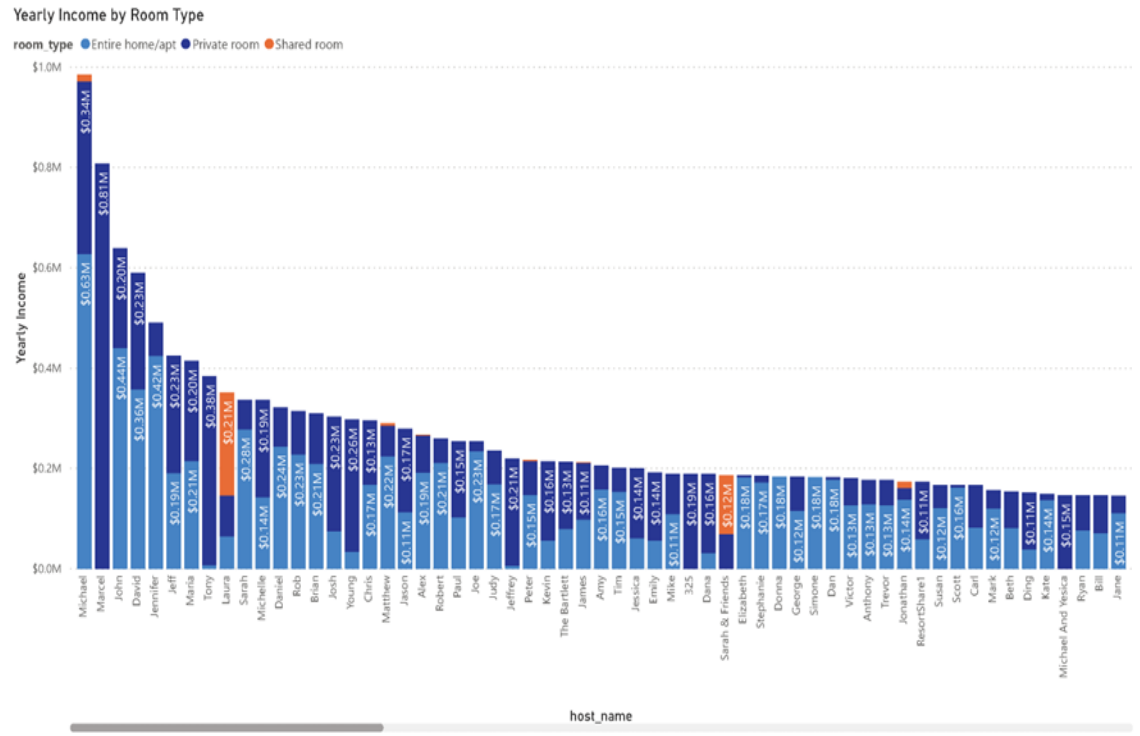
6. Pate, N. (2020). How Airbnb Uses Data Science to Improve Their Product and Marketing. Retrieved from https://neilpatel.com/blog/how-airbnb-uses-data-science/

7. Carrillo, G. (2019). Predicting Airbnb prices with machine learning and location data. Retrieved from https://towardsdatascience.com/predicting-airbnb-prices-with-machine-learning-and-location-data-5c1e033d0a5a


Average Price by Neighbourhood
neighbourhood
Fisherman"s Wharf
$613
$0
$200
$400
$600
Average of price
Instant Book by Reviews
instant_bookable False True
reviews_per_month
host_id
Listings Count
1
11.72
241
Room Type
Entire home/apt
Private room
Neighbourhood
Duboce Triangle
Excelsior
Financial District
Fisherman"s Wharf
Fisherman""s Wharf
Forest Hill
Glen Park
Haight-Ashbury
Value by Host
host_id
4966974 $41,459
20369590 $21,400
5182788 $18,000
28651428 $15,000
218110608 $13,890
39729330 $8,800
56042290 $8,792
76778950 $8,325
1504399 $6,440
115416167 $5,460
5708168 $3,874
188731766 $3,720
21376860 $3,237
$0K
$20K
$40K
$60K
Yearly Income
FISHERMAN'S WHARF
NORTH WATERFRONT
North Point Public Housing
RUSSIAN HILL
NORTH BEAC
Telegraph Hill
Bing


Average Price Per Year
Average of price
$350
$300
$250
$200
$150
$100
$50
$0
$175.00
$106.00
$275.00
$145.10
$298.47
$311.94
$265.57
$221.91
$227.70
$203.89
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
Year


Reviews Per Month by Income
reviews_per_month
Yearly Income
reviews_per_month
host_name
Michael
Michelle
David
Zarinah
John
Jessica
Maria
Jeff
Tony
Sarah & Friends
Jennifer
Sarah
Sonder (San Francisc...
The Bartlett
325
Mary Gaineng
Alex
Josh
Jason
Daniel


Yearly Income by Room Type
room_type
Entire home/apt
Private room
Shared room
Yearly Income
host_name