Customer Reviews are Associated with the Price of Airbnb Listings*

Kyunghyun Kim

April 27th, 2022

Abstract

Airbnb is an online home-sharing platform that connects people who want to rent out their homes with people who are looking for accommodations while they are traveling. Due to the outbreak of Covid-19, a lot of Airbnb hosts are experiencing a drop in revenue. This paper explores the various factors that impact the price of the Airbnb listings and builds a model that predicts how customer review rating scores impact the price of the listings. Although decisive causality could not be found, this paper discovered there was some relationship between customer reviews and pricing.

1. Introduction

Airbnb which stands for Air Bed and Breakfast is founded in 2008 and growing explosively by achieving unparalleled growth since it began. It gained a lot of popularity as an alternative to costly hotels. However, along with its rapid growth, there are emerging controversies about its service among many customers. As a community-based online rental platform, there are unavoidable negative consequences. The debate over which is better, hotels or Airbnb, continues to be a controversial subject among tourists and industry experts. Although Airbnb remains the dominant platform among online rental platform users, increased competition is causing its share to decrease. Some negative factors that lead to these negative consequences include communication, check-in service, cancellation policy, safety, location, and price. Among these factors, people complain most about the price of Airbnb. Many customers found that hotels are not necessarily more expensive than Airbnb as customers have to pay Airbnb's guest service fee which can be as much as 20% of the booking total. This Airbnb guest service fee includes a nightly rate, cleaning fee, and additional guest fee while hotels don't have these types of additional charges. The outbreak of Covid-19 has brought even more pressure on Airbnb hosts. According to the survey conducted by IPX 1031 (Lane, 2021), 70% of guests are fearful to stay at an Airbnb and 64% of guests either have cancelled or plan to cancel an Airbnb booking since the pandemic started. Due to Covid-19, Airbnb hosts are experiencing a 44% decrease in their revenue and lost $4036 since the pandemic began to spread in the US. This paper will focus on the factors that influence the price of the Airbnb listings and analyze and predict the pattern to help hosts to maximize their listings advantages and attractiveness to guests.

2. Data

2.1 Dataset

To attempt an analysis of factors that influence the price of the listings at Airbnb, I will primarily be examining Inside Airbnb data that was listed before December 05, 2021. The Inside Airbnb offers various Airbnb listings data from all around the world including the United States, Canada, and countries in Europe and Asia. I was interested in exploring Airbnb data in New York, one of the largest cities in the world. As one of the most famous cities in the world, New York place one of the largest economies in the world as it had a GDP of $1.7 trillion in 2020 (Ross, 2022). The diverse culture and myriad of entertainment options made New York to be a famous place for tourists. Welcoming over 66.6 million visitors annually, New York is the

*Code and data are available at: https://github.com/KyunghyunKim1224/final_paper.git

leading tourism destination in the world. The original data includes 38,277 Airbnb postings and 74 attributes including information about the neighbourhood, property type, room type, price, and guest reviews. Using R (R Core Team 2020), tidyverse (Wickham et al. 2019), ggplot2 (Wickham 2016), and dplyr (Wickham et al. 2021), I cleaned and extracted the necessary data to complete an exploratory analysis and modelling.

2.2 Data Cleaning

To begin, I start by removing all entries where the price is equal to $0 and higher than $350. Then, I removed the column that contains unnecessary information to the purpose of the paper such as URL, scrape id, and host verification. From the dataset, this paper focuses on the following: neighbourhood, borough, type of room and property and lastly customer review score on various factors.

2.3 Data Visualization

The first graph shows the number of listings by borough in NYC and by the room types. There are total of four types of listings: Entire home/apartment, Hotel room, Private room, Shared room. In most of the neighbourhood, it is clear that renting Entire home/apartment or private room out is most common. The second bar graph explores the average price of the listings by borough and the type of room. The average price of the listings is high in Manhattan and Brooklyn.

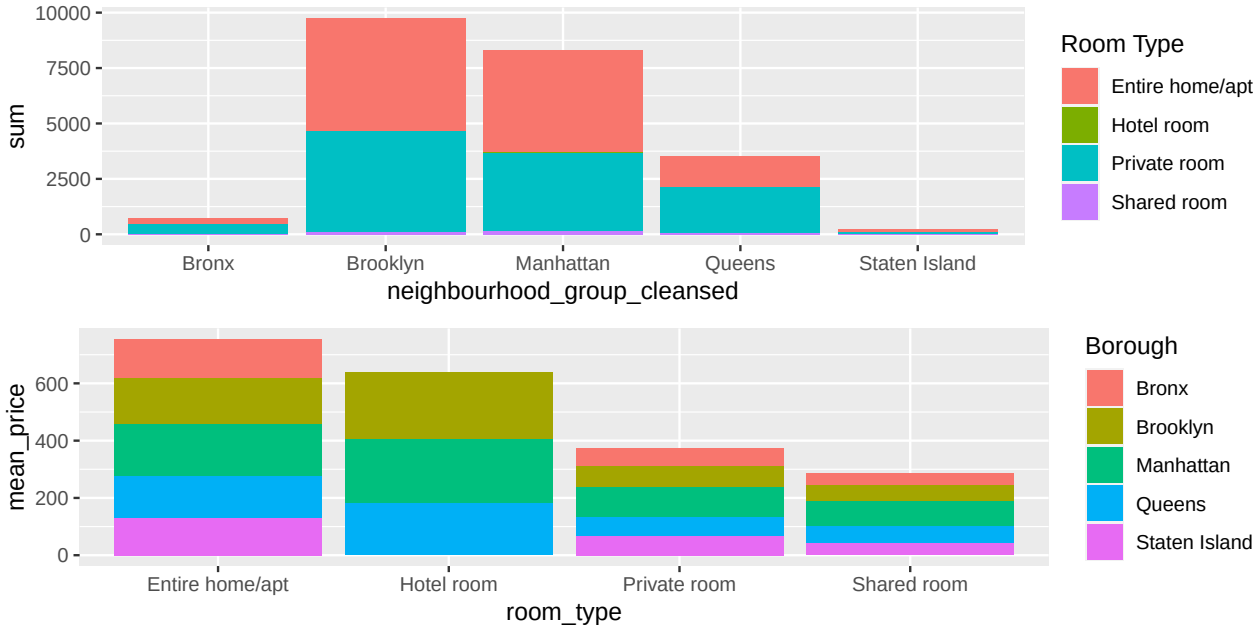

Figure 1: Average Price of Listings by Room Type and Borough

Below boxplots show the distribution of customer rating scores on different areas including accuracy - how accurately did the listing page represent the space?; cleanliness - did the guests feel that the space was clean and tidy?; check-in - did the guest feel that check-in process is easy to follow?; communication - how well did the host communicate with the guest before and during their stay?; location - how convenient they think the location is?; value - did the guest feel the listing provided good value for the price? The average customer rating scores on cleanliness is lower than other area as the mean is around 4.6.

3. Model

Since this paper aims to find whether customer review affects the price of the listings on Airbnb, the model it relies upon is multiple linear regression using R (R Core Team 2020). This model was appropriate for an

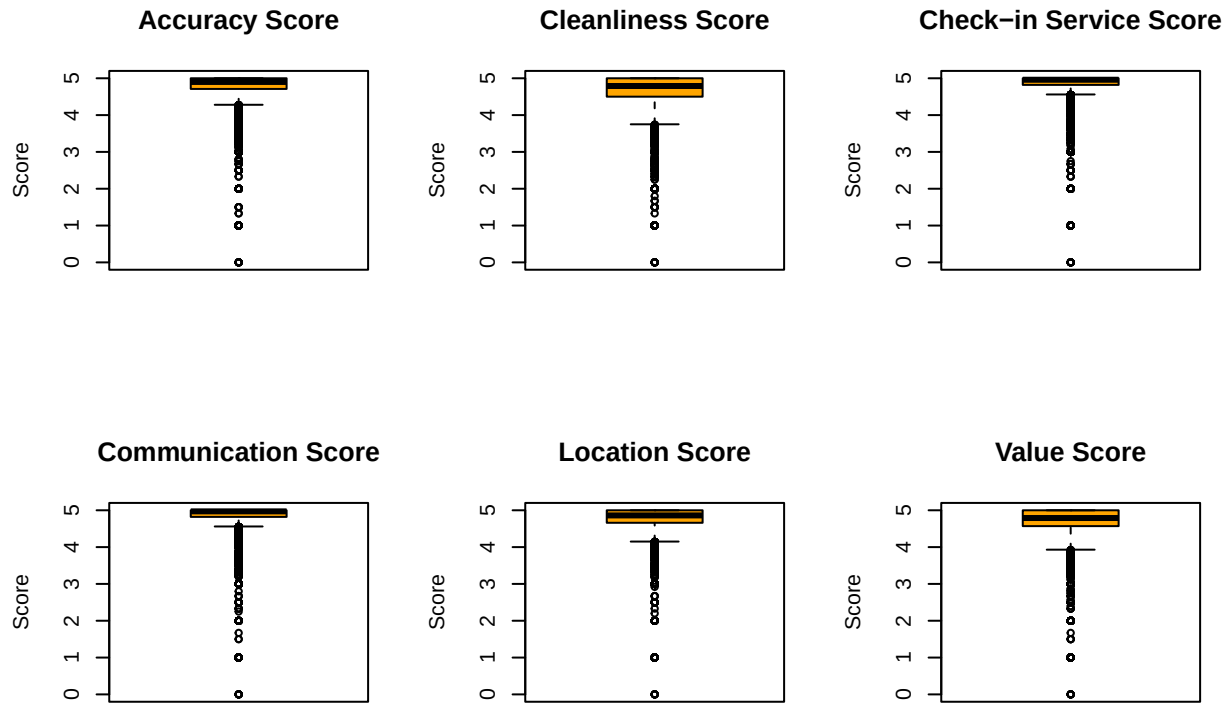

Figure 2: Boxplot of Customer Rating Scores on Various Areas

exploratory analysis of quantitative data, allowing for further investigation into the relationship and how it may be conducted in the future. The logistic regression is not nescessary in this paper since the dummy or binary variables are not found. This multiple liner regression will help explore any relationship that exists between customer review scores and the price of Airbnb listings. The model assumes that:

$$Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8$$

3.1 Selecting Predictors

In the first model, Y_1 is the price of the listing, X1 is the total number of guests reviews of each listings in NYC, X2 is the scores of guest review rating, X3 is the scores of accuracy, X4 is the scores of cleanliness, X5 is the score of check-in service, X6 is the scores of communication, X7 is the scores of location, and X8 is the scores of the value of the listings.
β_0 represents the predicted value of Y when X is 0 and β_1 to β_2 are the expected change to Y when X1 and X2 increase.

Next, I perform a stepwise process of eliminating variables that do not offer statistically significant value to the model. For selecting predictors, we will use the method of backward selection. It starts with our full model that includes all of the predictors, iteratively removes the least contributive predictors, and stops when we have a model where all predictors are statistically significant.

Model	DF	Sum of Squares	Residual Sum of Squares	AIC
- number_of_reviews	1	2	118035930	193082
- review_scores_communication	1	374	118036303	193082

The full model that includes all of the predictors has the AIC value of 193083.6. After removing the variable number_of_reviews and review_scores_communication, the AIC drops to 193082. Looking at below ANOVA table, we can confirm that all of our predictors are statistically significant as the p-value is all smaller than 0.05.

Predictors	Df	Sum Sq	Mean Sq	F value	Pr(>F)
review_scores_rating	1	848624	848624	162.045	< 2.2e-16 ***
review_scores_accuracy	1	72180	72180	13.783	0.0002057 ***
review_scores_cleanliness	1	1059231	1059231	202.260	< 2.2e-16 ***
review_scores_checkin	1	40450	40450	7.724	0.0054536 **
review_scores_location	1	1308248	1308248	249.810	< 2.2e-16 ***
review_scores_value	1	1325969	1325969	253.194	< 2.2e-16 ***

3.2 Checking Assuptions

Our final model reduces to:

$Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6$

Y_1 is the price of the listing, X1 is the scores of guest review rating, X2 is the scores of accuracy, X3 is the scores of cleanliness, X4 is the score of check-in service, X5 is the scores of location, and X6 is the scores of the value of the listings. With this final model, I first check whether or not the assumptions hold true. First, we check the constant variance assumption is hold as the homoscedasticity assumption is fulfilled. Then, we check normality of residuals is hold since the residuals are roughly normally distributed.

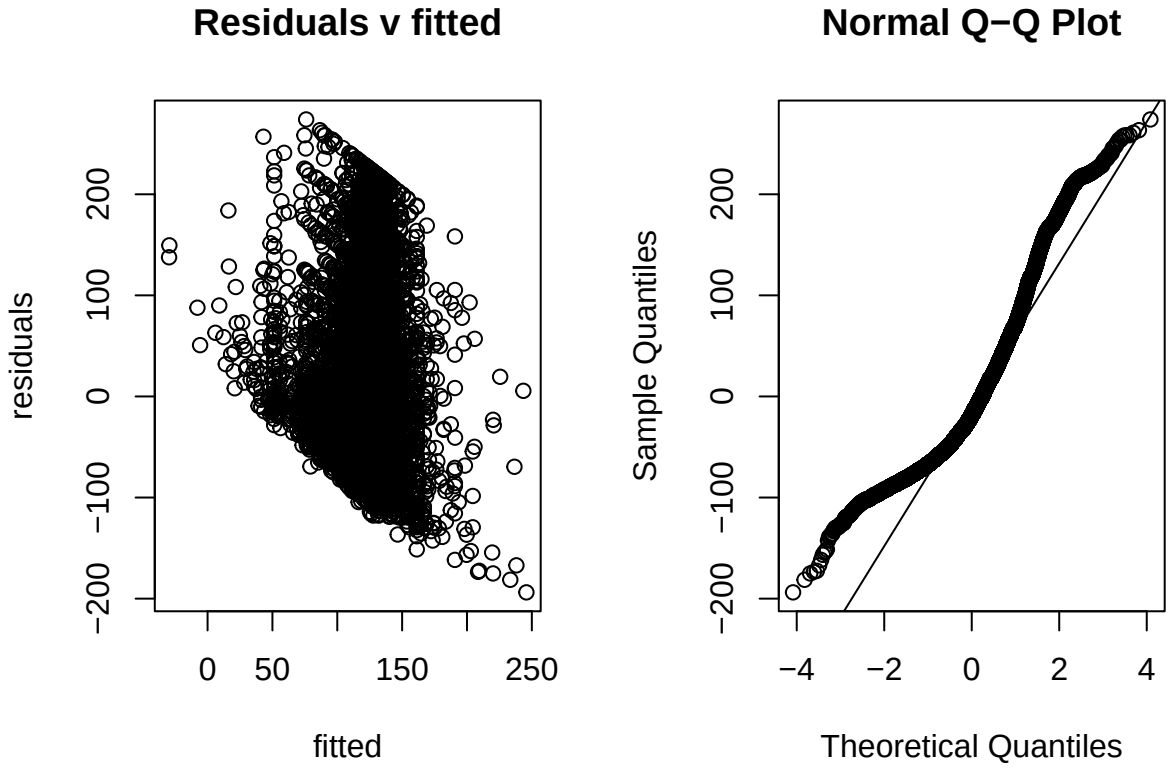

Figure 3: Model Assumptoins

4. Results

4.1 Model Diagnostics

Regression diagnostics are used to evaluate the model assumptions and investigate whether or not there are observations with a large, undue influence on the analysis. The first plot depicts residuals versus fitted values. I used this plot to check the assumption of linearity and homoscedasticity. Since the scatterplot of residuals do not have any distinctive patterns and the red line through the scatterplot is also almost straight, the linearity assumption is satisfied. From the QQ-plot, we observed each observation roughly falls on the straight line and there is no significant curve or break in the data. Therefore, we can assume normality. Then I created a scale-location plot to check the assumption of homoscedasticity. Since the red line on the plot is almost flat with equally and randomly spread data points, the homoscedasticity assumption is satisfied. In our fourth plot, we checked if the leverage of certain observations are driving abnormal residual distributions. There is no high leverage point in the data.

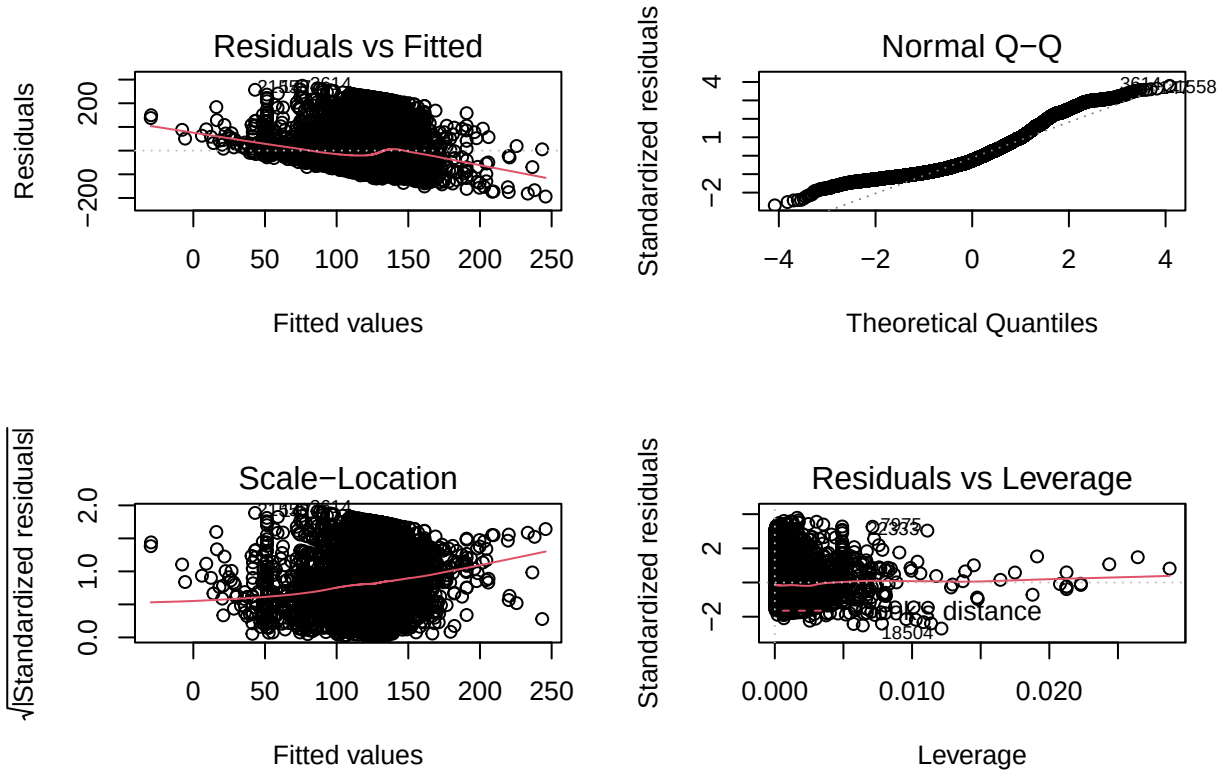

Figure 4: Model Diagnostics

4.2 Validation

After building our final model, I will determine the accuracy of this model on predicting the outcome for new unseen observations not used to build the model. In this step, we will estimate the prediction error. To do so, I first split my data into test and training sets so that 80% is used for training a linear regression model and 20% is used to evaluate the model performance. Then, calculated the Root Mean Squared Error (RMSE) which measures the average prediction error made by the model in predicting the outcome for an observation using Metrics (Hamner and Frasco 2018).

Comparing the RMSE of test dataset which is equal to 72.33103, and the RMSE of the full data which is 72.35012, we can see that the values are almost similar. Hence the variables that we have selected for our model are good estimators of our dependent variable.

5. Discussion

This paper explored how the price of NYC Airbnb listings varies by the customer rating review scores. By building a multivariable regression model, we obtained a model that predicts the price of the listings based on the customer ratings. In order to increase the price of the rent, hosts in NYC should take cleanliness as their top priority, and also accurately represent the rental space on the listing page. This model will help NYC Airbnb hosts to provide a better experience for guests while maximizing revenues. However, there are some limitations in this report. Although the model suggests that there is some relationship between the guests' rating scores and the price of listings, it is not a representation of a causal relationship. Other important factors could lead to high Airbnb listing prices in NYC. For instance, the data that is used in this paper is from December of 2021. Therefore, it cannot represent the whole rental market in NYC as the patterns and findings discovered throughout this paper are limited to the specific year from which the data is provided.

Appendix

Datasheet for Dataset

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
   - The dataset was created to enable analysis of Australian politicians. We were unable to find a publicly available dataset in a structured format that had the biographical and political information on Australian politicians that was needed for modelling.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
   - Kyunghyun Kim
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
   - N/A
4. *Any other comments?*
   - N/A

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
   - Each row of the main dataset is Airbnb listing, and contains the information about that specific listing such as location, host, price, amenities, guest rating scores.
2. *How many instances are there in total (of each type, if appropriate)?*
   - The original data contains 38277 instances and this paper analyzed 28087 instances.
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
   - It is a sample. The larger set is the original data that contains 38277 listings. The sample is the representative of the larger set. It was validated by comparing the values of the root mean squared error of testing and training data.
4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*
   - The raw data consists of demographic, descriptive information about listings and quantitative information about the listings.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
   - No
6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*
   - There are some missing information in the raw data and this is due every listings have different features.
7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*
   - No
8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*
   - Yes. The data was split into test and training sets so that 80% is used for training a linear regression model and 20% is used to evaluate the model performance.
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

- No

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*
    - The data rely on external resources, Airbnb app.
11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*
    - No
12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
    - No
13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*
    - No
14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*
    - Yes. The name and contact informatoin about hosts are recorded in the data.
15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*
    - Yes, it contains the detailed information about Airbnb hosts.
16. *Any other comments?*
    - N/A

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects, or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*
    - The data was directly observable.
2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*
    - The data was collected through the software programs, Airbnb app.
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
    - N/A
4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*
    - People who are renting out their properties on Airbnb.
5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*
    - The data was collected throughout December 2021.
6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

- No

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
   - Data was obtained via third parties, Inside Airbnb.
8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*
   - Yes, they were told that their answers would be recorded.
9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*
   - Yes. The detailed listings data include the information about the hosts and if they refuse to answer, their answers are not recorded.
10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*
    - No
11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - No
12. *Any other comments?*
    - No

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*
   - Yes, the data was cleaned and missing values are removed.
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*
   - No
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
   - No
4. *Any other comments?*
   - No

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
   - N/A
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
   - No
3. *What (other) tasks could the dataset be used for?*
   - It could also be used for predicting the customer rating scores based on the host performance.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups or other risks or harms? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
   - N/A
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
   - No, none were stated.

6. *Any other comments?*
    - N/A

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
    - N/A
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - N/A
3. *When will the dataset be distributed?*
    - N/A
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
    - N/A
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
    - N/A
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
    - N/A
7. *Any other comments?*
    - N/A

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - Kyunghyun Kim
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - kyunghyun.kim@mail.utoronto.ca
3. *Is there an erratum? If so, please provide a link or other access point.*
    - No
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
    - No
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
    - N/A
6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*
    - Yes. The older version of the dataset is available in the Inside Airbnb website.
7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*
    - N/A
8. *Any other comments?*
    - N/A

References

Airbnb. (2021, December). Get the Data. Inside Airbnb. Retrieved April 27, 2022, from http://insideairbnb.com/get-the-data/

Auguie, Baptiste. 2017. gridExtra: Miscellaneous Functions for "Grid" Graphics. https://CRAN.R-project.org/package=gridExtra.

Guttentag, D. (2019). Progress on airbnb: A literature review. Journal of Hospitality and Tourism Technology, 10(4), 814–844. https://doi.org/10.1108/jhtt-08-2018-0075

Hamner, Ben, and Michael Frasco. 2018. Metrics: Evaluation Metrics for Machine Learning. https://CRAN.R-project.org/package=Metrics.

Hati, S. R., Balqiah, T. E., Hananto, A., & Yuliati, E. (2021). A decade of systematic literature review on airbnb: The sharing economy from a multiple stakeholder perspective. Heliyon, 7(10). https://doi.org/10.1016/j.heliyon.2021.e08222

Lane, L. (2021, June 28). How bad are covid-19 pandemic effects on airbnb guests, hosts? Forbes. Retrieved April 27, 2022, from https://www.forbes.com/sites/lealane/2020/06/09/how-bad-are-covid-19-pandemic-effects-on-airbnb-guests-hosts/?sh=48c449027432

R Core Team. 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ross, S. (2022, February 8). New York's economy: The 6 industries driving GDP growth. Investopedia. Retrieved April 27, 2022, from https://www.investopedia.com/articles/investing/011516/new-yorks-economy-6-industries-driving-gdp-growth.asp

Tian, Z. (2021). Use python data analysis to gain insights from Airbnb hosts. Advances in Mathematical Physics, 2021, 1–10. https://doi.org/10.1155/2021/1079850

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. Dplyr: A Grammar of Data Manipulation. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley. 2016. Ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. https://ggplot2.tidyverse.org. ———. 2020. Httr: Tools for Working with URLs and HTTP. https://CRAN.R-project.org/package=httr. ———. 2021. Rvest: Easily Harvest (Scrape) Web Pages. https://CRAN.R-project.org/package=rvest.

Xie, Yihui. 2021. Knitr: A General-Purpose Package for Dynamic Report Generation in r. https://yihui.org/knitr/.

Zhu, Hao. 2020. kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax. https://CRAN.R-project.org/package=kableExtra.