Analysis of Airbnb Price Determinants Using Linear Regression Model

December 4, 2023

Names of Group Members	Contribution to Proposal
Yuxuan Wu	Methods; Reference
Xinrui Zhai	Results-code
Xile Chen	Results-analysis; Discussion
Junsoo Kim	Introduction

1.Introduction

Our study delves into the intricate dynamics of Airbnb pricing. It focuses on a central question: "What are the main factors influencing Airbnb room rates, and how do they quantitatively impact pricing?" We aim to identify and analyze variables significantly influencing room rates, encompassing aspects such as property type, room type, location, and accommodation features. Using a linear regression model, our analysis quantifies the impact of these variables on room rates, offering insights into the attributes most profoundly affecting pricing decisions.

Part 1 of our project includes a comprehensive literature review, focusing on three peer-reviewed articles that examine themes within the Airbnb market. These studies provide a foundation but highlight gaps, particularly in the comprehensive analysis of diverse variables across multiple locations. Our research seeks to fill these gaps by incorporating a broader range of variables, thus enhancing the existing knowledge base. Our methodology, grounded in academic literature and a thorough examination of the Airbnb rental platform, aims to deepen understanding of Airbnb pricing dynamics, aiding hosts in refining their pricing strategies and assisting guests in making informed accommodation choices.

Chattopadhyay and Mitra's 2019 research is foundational to our study, highlighting the significance of various variables in Airbnb room pricing. They emphasize the importance of amenities and other listing attributes in shaping room rates, with their analysis covering 11 US cities. Another pivotal study is on geographic variation in Airbnb pricing, focusing on Shanghai's market. It underscores the role of property quality in listing prices and explores the impact of location conditions, particularly in areas with robust transportation networks. The study employs various linear and geographically weighted regression models to analyze price determinants.

A third essential reference is a 2015 motivation-based segmentation study on Airbnb tourists. This research investigates why tourists choose Airbnb accommodations, identifying five motivating factors: Interaction, Home Benefits, Novelty, Sharing Economy Ethos, and Local Authenticity. This segmentation offers valuable insights into consumer preferences and behaviours in the Airbnb market.

These references collectively shape our research approach, providing a comprehensive backdrop for examining Airbnb pricing dynamics and the factors influencing room rates. Our study aims to offer a nuanced understanding of Airbnb's pricing mechanisms, benefiting both hosts and guests. By integrating diverse viewpoints, our research contributes valuable insights to scholarly discussions and practical applications within the Airbnb community.

2. Methods

2.1 Dataset Description

Our study utilized data from an open-source dataset available at data. World, explicitly focusing on Airbnb listings in Georgia (URL: https://data.world/cannata/gaairbnb). This rich dataset encompasses various variables pertinent to Airbnb listings, such as Property Type, Room Type, Zipcode, Accommodation Capacity, Number of Bathrooms, Bedrooms, Beds, Overall Rating Score, Host Response Time, and Minimum Nights. These variables were carefully chosen based on their potential influence on Airbnb pricing, aligning with our research objective to identify factors affecting Airbnb rates. The dataset provides a comprehensive view of Airbnb's market dynamics in Georgia, offering a robust foundation for our statistical analysis.

2.2 Exploratory Data Analysis (EDA)

A crucial part of our methodology was conducting rigorous exploratory data analysis (EDA). We meticulously selected and classified variables into numerical and categorical types, including Property Type, Room Type, Zipcode, and Host Response Time. Data cleansing was crucial, so we created dummy variables for categorical data. We used scatter plots and numerical summaries to understand data distribution and relationships between variables, revealing important data patterns.

2.3 Model Fitting and Transformation

Fitting the model was a central part of our methodology. We defined our linear model as Price ~ Property Type + Room Type + Zipcode + Accommodation Capacity + Number of Bathrooms + Number of Bedrooms + Number of Beds + Rating Score + Host Response Time + Minimum Nights. We conducted the residual analysis to ensure model robustness, looking at residual plots and applying transformations like Box-Cox to achieve residual normality, as shown in QQ plots. We thoroughly assessed and transformed the model, checking for regression assumption violations such as independence, constant variance, and non-linearity. We made appropriate adjustments, like adding polynomials or interaction terms, and used F-tests and 95% confidence intervals to assess the overall model significance.

2.4 Statistical Analysis

We conducted a comprehensive statistical analysis, examining R-squared and Adjusted R-squared to evaluate the model's explanatory power. We consulted the ANOVA table to understand the variance within and between categorical variable groups. We determined p-values for each predictor to test the null hypothesis that the coefficient is zero and conducted T-tests for each predictor to validate our findings further.

2.5 Conclusion

In concluding our methodology, we closely examined each predictor's coefficient estimate, assessing its statistical significance with t-values and p-values. This led to a discussion on the model's fit and the broader implications of our results for the variables of interest. Our rigorous methodology was crucial in ensuring the reliability and validity of our findings, adding meaningful insights to the study of factors influencing Airbnb rates.

3. Results

3.1 Description of Data

Table 3.1: Summary the data of the numerical variables for train and test datasets

Variables	**Min (Train)**	**Min (Test)**	**1st Qu. (Train)**	**1st Qu. (Test)**	**Median (Train)**	**Median (Test)**	**Mean (Train)**	**Mean (Test)**	**3rd Qu. (Train)**	**3rd Qu. (Test)**	**Max (Train)**	**Max (Test)**
Accommodation Capacity	1	1	2	2	2	2	3.140	3.125	4	4	16	16.0
Number of Bathrooms	0	0	1	1	1	1	1.100	1.106	1	1	4	7.5

Variables	Min (Train)	Min (Test)	1st Qu. (Train)	1st Qu. (Test)	Median (Train)	Median (Test)	Mean (Train)	Mean (Test)	3rd Qu. (Train)	3rd Qu. (Test)	Max (Train)	Max (Test)
Number of Bedrooms	0	0	1	1	1	1	1.424	1.396	2	2	10	10.0
Number of Beds	1	1	1	1	1	1	1.985	1.947	2	2	16	16.0
Rating Score	45	40	91	91	95	95	93.700	93.790	99	98	100	100.0
Minimum nights	30	19	165	170	250	255	339.800	334.300	414	400	4500	3500.0

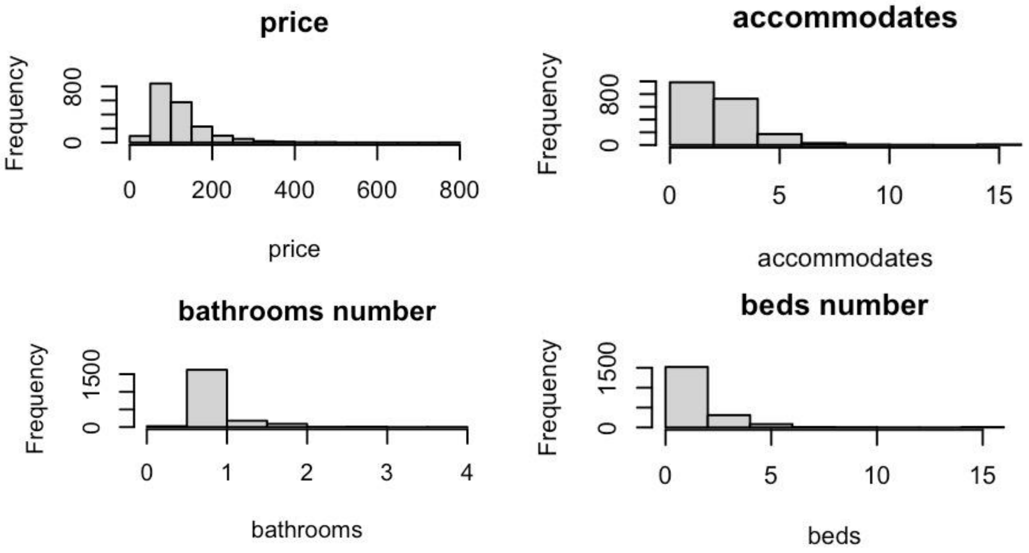

Figure 3.1.1: Histograms of some numerical variables

The above histograms indicate data distribution of observations in some numerical variables, the variables of price and number of bathrooms are the most normal distributed.

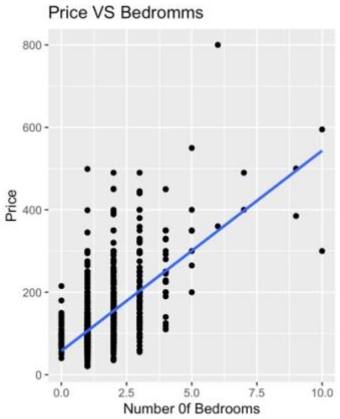

Figure 3.1.2: Scatter plots of numerical variable Number of Bedrooms

The scatter plots indicate the relationship between the predicator and the response, which means the overall trend of the plot is going up.

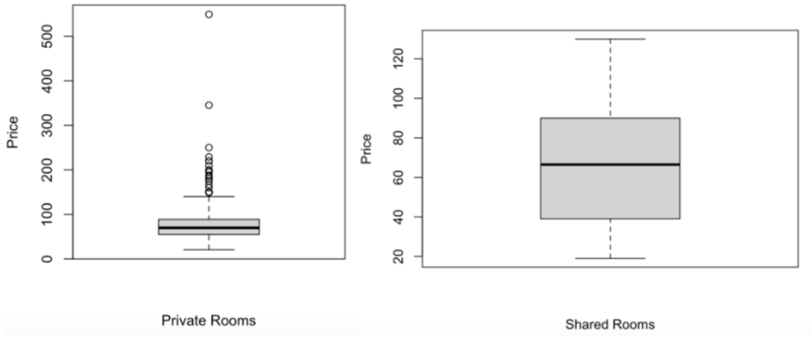

Figure 3.1.3: Boxplots of categorical variable room type with private room and shared room

The boxplots show the overall patterns of prices for each category of private room and shared room of categorical variable room type.

3.2 Process of Obtaining Final Model

Step 1: Based on our interest and literature information about Airbnb price to choose variables, and consider the corresponding graphs, form Model 1as follows:

$\widehat{price}$ = 25.0027 + 7.0826 * I(Property Type = Bed & Breakfast) + 13.7120 * I(Property Type = Boat) + 16.4798 * I(Property Type = Cabin) - 64.6435 * I(Property Type = Camper/RV) + 13.4736 * I(Property Type = House) - 1.6102 * I(Property Type = Loft) - 4.8332 * I(Property Type = Other) + 13.0746 * I(Property Type = Villa) - 34.2382 * I(Room Type = Private room) - 36.3983 * I(Room Type = Shared room) - 39.1717 * I(Zipcode in East) - 46.6595 * I(Zipcode in North) - 18.5502 * I(Zipcode in South) - 29.8504 * I(Zipcode in West) + 9.7789 * Accommodation Capacity + 29.4228 * Number of Bathrooms + 21.8485 * Number of Bedrooms + 1.5569 * Number of Beds + 0.2468 * Rating Score - 0.9627 * I(Host Response Time = N/A) + 7.9441 * I(Host Response Time = Within a day) + 8.4760 * I(Host Response Time = Within a few hours) + 1.4117 * I(Host Response Time = Within an hour) - 0.7169 * Minimum Nights

Step 2: Choose the most significant variables with 2 stars or 3 stars on the right side of the list from Model 1 to form Model 2.

Step 3: Using stepwise selection to choose the best model from Model 2 until AIC gets bigger no matter how you change it to form Final Model (Model 3).

$\widehat{price}$ = 49.056 - 32.166 * I(Room Type = Private room) - 36.544 * I(Room Type = Shared room) - 40.046* I(Zipcode in East) - 45.395* I(Zipcode in North) – 19.754 * I(Zipcode in South) – 31.346 * I(Zipcode in West) + 11.003 * Accommodation Capacity + 30.813 * Number of Bathrooms + 23.742 * Number of Bedrooms

3.3 Goodness of Final Model (Model 3)

(1) Check if the model satisfies condition 1 and condition 2

Condition 1: Condition 2:

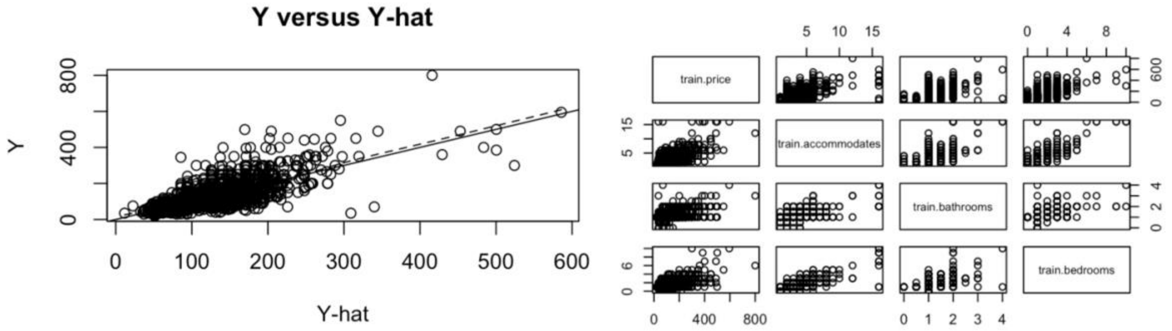

Figure 3.3.1: Plots for checking the satisfaction in condition 1 & 2

In the left plot which indicates the response versus fitted values, the points near or on the line, then condition 1 is satisfied. In the right plot which indicate relations, the numerical variables are almost linear, then condition 2 is satisfied.

(2) Check the assumptions in regression by residual plot and QQ plot

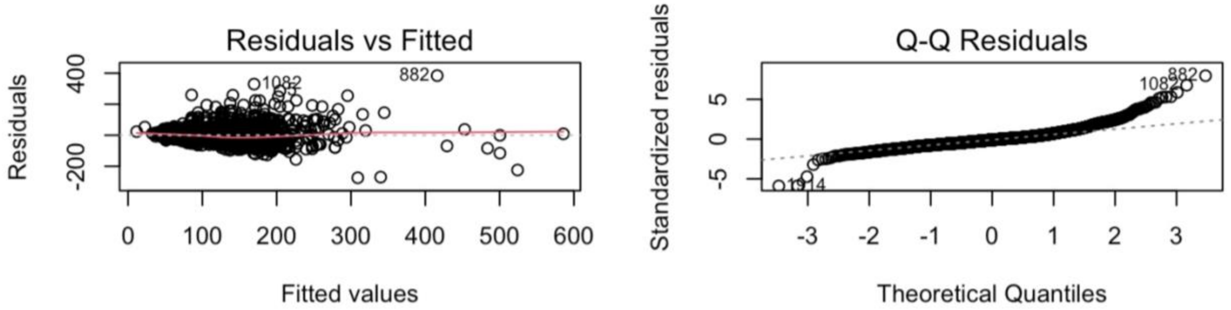

Figure 3.3.2: Residual Plot and QQ Plot

In the residual plot, there is no systematic pattern, cluster pattern, or fanning pattern. In the QQ plot, all the points are near or on the straight line and there are very few points of deviation. Then, the final model satisfies the 4 assumptions in regression.

(3) Check if there is any problematic observations

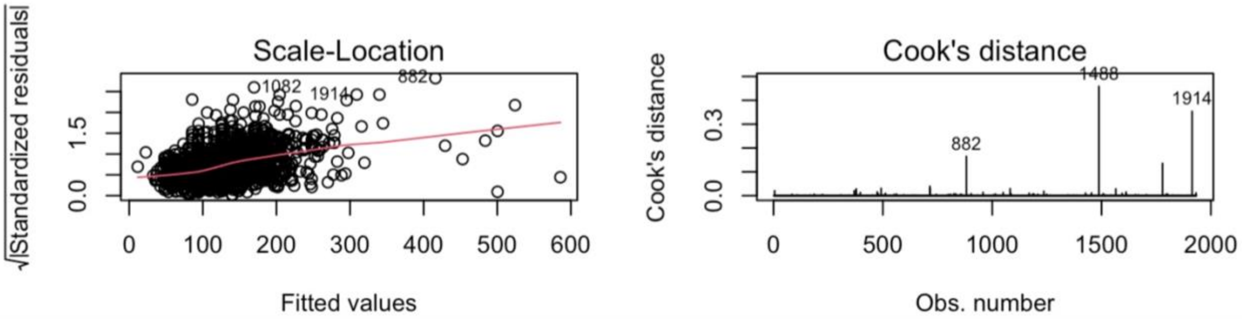

Figure 3.3.3: Scale-Location Plot and Cook's Distance

There are 15 outliner points, 151 leverage points and 17 Beta 0; 6 Beta 1; 5 Beta 2; 17 Beta 3; 16 Beta 4; 17 Beta 5 influential points, which are quite few based on the size of the dataset.

(4) Check the final model most preferred by partial F test

Create a new model which includes some but not all variables in the final model randomly. By using ANOVA formula to compare two models, we have P-value less than 0.05, which means the final model better.

(5) Check model by testing dataset

Table 3.3 The summery of 3 models

Models	RSS	Rsquare	Rsquare_adj	AIC	AIC_c	BIC
Model 1	4,496,568	0.5300041	0.5240953	15,039.36	15,040.10	15,188.11
Model 3	4,589,054	0.5203372	0.5180935	15,048.74	15,048.87	15,113.98
Test Model	4,589,054	0.5203372	0.5180935	15,048.74	15,048.87	15,113.98

The test model is consistent with Model 3, which shows the goodness of the final model.

4. Discussion

4.1 Conclusion

The final model is as follows:

$\widehat{price}$ = 49.056 - 32.166 * I(Room Type = Private room) - 36.544 * I(Room Type = Shared room) - 40.046* I(Zipcode in East) - 45.395* I(Zipcode in North) – 19.754 * I(Zipcode in South) – 31.346 * I(Zipcode in West) + 11.003 * Accommodation Capacity + 30.813 * Number of Bathrooms + 23.742 * Number of Bedrooms

From the final model, consider the numerical variables and categorical separately,

(1) 3 Numerical Variables:

We can conclude that there is a positive relationship between price and accommodation capacity given all other variables constant since the coefficient is positive (11.003), which means as the accommodation capacity increase, the price increase. Similarly, the other two numerical variables (Number of Bathrooms, Number of Bedrooms) have positive relationship with price.

This results align with intuitive expectations and is consistent with existing literature, which suggests that properties with more space and more private or additional facilities with suitable upper bound tend to command higher prices in the real estate market.

(2) 2 Categorical Variables:

With all coefficients associated with the indicators of a categorical variable are negative for both Room Type and Zipcode, it typically suggests that the reference category (the category not expressed by indicators) is the preferred or most commonly chosen category among the respondents or observations.

For room type, this means the entire home/apt is most preferred room type, which is consistent that guests often gauge their willingness to pay based on the level of privacy they will receive.

For Zipcode, this means Central area is most preferred in Amsterdam. By analysing the map of Amsterdam, this is the city centre, where business centres and cultural attractions are concentrated, and with convenient transportation, there is a high demand for housing in this area, especially hotels and Airbnb house for tourists, leading to an increase in prices. This is constent with Jiang and Zhang’s study (Jiang et al., 2022).

4.2 Limitation

The final model has some limitations. First, the relatively small dataset may restrict the study's ability to capture all relevant patterns comprehensively. In addition, we can explore more advanced models that can enhance predictive capabilities and result in a more precise and comprehensive analysis through model comparison. Also, we contain some problematic observations (shown in 3.3(3)). We kept them to ensure the accuracy of analysis and integrity of the dataset; however, this may introduce noise to the results.

Reference

Chattopadhyay, M., & Mitra, S. K. (2019). Do airbnb host listing attributes influence room pricing homogenously? *International Journal of Hospitality Management*, *81*, 54–64. https://doi.org/10.1016/j.ijhm.2019.03.008

Guttentag, D., Smith, S., Potwarka, L., & Havitz, M. (2017). Why tourists choose airbnb: A motivation-based segmentation study. *Journal of Travel Research*, *57*(3), 342–359. https://doi.org/10.1177/0047287517696980

Jiang, Y., Zhang, H., Cao, X., Wei, G., & Yang, Y. (2022). How to better incorporate geographic variation in airbnb price modeling? *Tourism Economics*, *29*(5), 1181–1203. https://doi.org/10.1177/13548166221097585

Wang, H. (2023). *Predicting Airbnb Listing Price with Different models - DRP*. DRPress. Retrieved from https://www.drpress.org/predicting-airbnb-listing-price-with-different-models

Airbnb. (n.d.). *Customized Regression Model for Airbnb Dynamic Pricing*. KDD. Retrieved from https://www.kdd.org/kdd2018/accepted-papers/view/customized-regression-model-for-airbnb-dynamic-pricing

Liu, P. (2021). *Airbnb Price Prediction with Sentiment Classification*. San Jose State University. Retrieved from https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=1979&context=etd_projects

Alharbi, Z. H. (2023). *A Sustainable Price Prediction Model for Airbnb Listings*. MDPI. Retrieved from https://www.mdpi.com/2071-1050/15/17/13159

McNeil, B. (2020). *Price Prediction in the Sharing Economy: A Case Study with Airbnb*. University of New Hampshire. Retrieved from https://scholars.unh.edu/cgi/viewcontent.cgi?article=1511&context=honors

Toader, V. (2022). *Analysis of price determinants in the case of Airbnb listings*. Taylor & Francis Online. Retrieved from https://www.tandfonline.com/doi/full/10.1080/1331677X.2021.1962380