Proceedings of ICFTBA 2025 Symposium: Data-Driven Decision Making in Business and Economics
DOI: 10.54254/2754-1169/2025.BL30924

A Study of Airbnb Rental Prices Based on Linear Regression Models

Xinyu Shuai

Kunming No.1 High School International Division, Kunming, China
sallyxinyushuai@outlook.com

Abstract. Finding the best rental prices for Airbnbs is very tricky because of all the differences between houses, how cancellations work, and different services, which makes it hard for owners to get the most money and for places like Airbnb to make things better. In order to solve the problem, this paper uses multiple Linear regression models to predict Airbnb rental prices using the publicly available Airbnb list, so as to achieve the goal of predicting performance and interpretability. Price normalization using logarithm, filling missing values with median, and handling of multicollinearity through the application of the VIF test and variable selection, are all part of data preprocessing. Experimental results show that the model performs solidly, with a Mean Squared Error (MSE) of 0.2896, Root Mean Squared Error (RMSE) of 0.5382, and Mean Absolute Percentage Error (MAPE) of 8.69%, indicating relatively low deviation in both absolute and relative terms. the model's R^2 value is as high as 0.4941 (adjusted R^2: 0.4937), which suggests that the model can explain about 49.4% of the variation in the logarithm of the rental price. This study has demonstrated the application of linear regression in quantifying price determinants in short-term rentals, which can provide transparent and actionable insights for both hosts to improve their listing settings and for platforms to improve the matching efficiency of users. Model's interpretability is also good at regulating because it can make clear how the pricing mechanism works, so this kind of model is quite helpful to improve a clearer and healthier short-term rental ecosystem.

Keywords: Airbnb price prediction, multiple linear regression model, short-term rental market

1. Introduction

Airbnb is an online platform that lets people rent out short-term accommodations, and the hosts on Airbnb span from people who have an extra bedroom to professional property managers with many listings. Listing features vary widely: rooms, accommodates, cancellations, cleaning fees [1]. It's a complicated challenge for both hosts and the platform to find an ideal price. Explore and visualize the data to get linear regression models of forecasting Airbnb rental price that give useful info and suggestions for making the hosts' money grow, guests happier, and the whole platform run better. This paper seeks to create such a model, and attempts to answer the following research questions:

- How good can a linear regression model be for the Airbnb datasets?

DOI: 10.54254/2754-1169/2025.BL30924

• How could these models improve our knowledge about what affects prices in short-term rental market?

Existing studies stress the importance of forecasting Airbnb rental prices. For example, Cai and Zhou analyzed the effects of 5 groups of explanatory variables on Airbnb price in Hong Kong [2]. Hwang and Gupta studied changes in neighborhoods and residential instability as it related to affordable housing [3]. Quattrone et al. reviewed the use of geographic and non-geographic variables in Airbnb price prediction [4]. But few studies have looked closely at what's good and not so good about linear regression models for this kind of work, and none have made easy-to-use, clear models that can do both predicting and explaining well.

To address this gap, a multiple linear regression model has been made to evaluate and predict the rent of an Airbnb. The model contains the variables of room_types, accomodates, cancellation_policy, cleaning_fee, bathrooms and instant_bookable [1]. The biggest advantage of Linear Regresstion, one has after using it is Interpretability, very easily you can see and understand how a single variable impacts the Price. This transparency is especially pertinent for the governance of platforms and regulatory oversight because it allows Airbnb to provide explanations and justifications for pricing systems. And such clarity can help to resolve outstanding problems that previous studies might have failed to address, because the models are too complicated for practical accountability [5].

2. Literature review

Recently related to the housing and short-term rental market analysis, the research showed that property price prediction is very important for understanding the market behavior, community changes and platform regulations. Three recent studies are most relevant for Airbnb housing price analysis, but with slightly different aims from the price prediction directly. They offer views of data structure, variable selection, and model complexity vs. interpretability.

The analysis and prediction of the Spatial Penetration of Airbnb in US cities (2018)

explored how Airbnb listings had spread out in major U. S. cities, finding out the main socioeconomic and geographic factors leading to this kind of spatial invasion. Authors use publicly available Airbnb listing data, U. S. Census 2010 Data, and cost of living index to develop spatial regression model that explains variations in Airbnb listing density. The findings show that population size, median income, age, and housing costs have an impact on the Airbnb activity. Although this study does not try to forecast the prices themselves, it shows that the spaces and social-economic aspects will be able to help understand how the Airbnb marketplaces are working well, which makes an awesome starting point on deciding which aspects I will want to test if they can help decide the housing prices. The model uses transparent regression methods so we can understand how the explaining variables are connected to listing penetration [4].

This was a study on long-term neighborhood change and residential instability in Oakland conducted by the Federal Reserve Bank of San Francisco and Stanford University's Changing Cities Research Lab. The dataset merged admin housing data, Equifax credit panel data and geographic info to calculate migration rate, credit score and overcrowding pattern. The analysis showed that people with low socioeconomic status (SES) have higher rates of moving around and having unstable housing, especially when they live in areas where things are getting nicer and fancier. While not directly related to either Airbnb or predicting price, these are the community level social economic features as well as a variable on the building turnover. These two variables will have a major impact in affecting the rental market. Neighborhood variables for Airbnb price prediction

DOI: 10.54254/2754-1169/2025.BL30924

could make the model more interpretable and fairer by giving the pricing pattern a place in the community structure [3].

Cal and Zhou's study look into other actions of the housing market. The details here are light so could just be using some sort of regression model or something that's looking at time-series models of factors influencing prices. The study shows a growing trend among housing researches toward incorporating economic measures and spatial information in modeling the value of homes. Though like so much modern work its goal is to predict, not explain [2], transparent models such as linear regression have research opportunities in short term rentals.

The reviewed studies together show how housing market prediction is becoming more important to understand changes in society, space, and economy. but these things do show there is an ongoing give and take about the likelihood that a model has guessed right and how easily we know why it made that guess. To tackle these shortcomings, this proposed research will build a Multiple Linear Regression Model to predict the Airbnb Listing Price. By incorporating socioeconomic, geographical, and platform-specific features, the objective is to improve the model's transparency and interpretability. And the model would generate not just good predictions for prices, but good sense that can be used by regulatory bodies at the Airbnb platform to craft good open markets.

3. Methodology

3.1. Data collection and description

The data of this research is open Airbnb list; it contains the specific information about houses and the host and also contain various price variables. Cleaning of the data has been carried out prior to the analysis is kept accurate and reliable. Firstly, the id column is dropped, as it is merely a marker that does not offer analytical value to the model. Secondly, duplicates were removed to avoid data redundancy and bias in the regression.

After cleaning, there were over 50,000 valid listings for the New York area with several key numerical features related to price prediction. As can be seen in Table 1, the number of people accommodated is on average 3.16, most of them can fit between 2 to 4 people, which we can tell from the 25th, 50th and 75th percentile being 2.0, 2.0 and 4.0. The mean number of bathrooms is 1.24 and the mean numbers of bedrooms and beds are around 1.27 and 1.71, respectively, suggesting that the data set contains mostly small-to-medium sized properties. review_scores_rating variable has a mean value of 94.05 and a standard deviation of 7.84 which indicates that the guest satisfaction is quite high for the majority of the listings and the minimum and maximum values of 20.0 and 100.0 respectively indicate the range of guest experiences. The dependent variable, act_price, had a mean of 160.37 with some variation (standard deviation = 168.58) and a large range of 1.0 to 1999.00, representing the various sizes, amenities, and locations of the properties.

Room_type is a categorical variable mostly "Entire home/apt" 41308 listings, cancellation_policy is mostly "strict" 32500 listing, cleaning_fee mostly "True" 54399 listings, and instant_bookable mostly "1" 54660 listings also give context to the property. In general, the dataset gives an all-around and representative basis for modeling Airbnb rental costs, both numerical and categorical characteristics providing multi-dimensional perspectives on listing traits.

Table 1. Statistical summary of the dataset

	count	unique	top	freq	mean	std	min	25%	50%	75%	max	describe	unit
room_type	74106	3	Entire home/apt	41308	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Type of Room in the property	category
accommodates	74108.0	NaN	NaN	NaN	3.155125	2.153603	1.0	2.0	2.0	4.0	16.0	How many adults can this property accommodate	adults
bathrooms	73908.0	NaN	NaN	NaN	1.235272	0.582054	0.0	1.0	1.0	1.0	8.0	Number of bathrooms on the property	units
cancellation_policy	74103	3	strict	32500	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Cancellation policy of the property	category
cleaning_fee	74107	2	True	54399	NaN	NaN	NaN	NaN	NaN	NaN	NaN	This denotes whether the property cleaning fee...	category
instant_bookable	74111	2	f	54660	NaN	NaN	NaN	NaN	NaN	NaN	NaN	It indicates whether an instant booking facile...	category
review_scores_rating	57389.0	NaN	NaN	NaN	94.067365	7.836556	20.0	92.0	96.0	100.0	100.0	Review rating score of the property	score
bedrooms	74019.0	NaN	NaN	NaN	1.265797	0.852149	0.0	1.0	1.0	1.0	10.0	Number of bedrooms in the property	units
beds	73980.0	NaN	NaN	NaN	1.710868	1.254142	0.0	1.0	1.0	2.0	18.0	Total number of beds in the property	units
act_price	74111.0	NaN	NaN	NaN	160.370849	168.580415	1.0	75.0	111.0	185.0	1999.000001	Actual rental price of the property for a fixe...	dollars

All of the missing values in a database for each variable are gathered. In which 0.26% missing values for bathrooms, 15.1% missing values for review_scores_rating, 0.12% missing values for bedrooms and 0.17% missing values for beds. room_type, accommodates, cancellation_policy, cleaning_fee, instant_bookable, act_price all have 0 missing values. The missing values for each variable after using the median to impute them is collected.

3.2. Data pre-processing

Before building the model, some preprocessing operations were carried out to make sure that the dataset was tidy, coherent, and appropriate for performing linear regression.

The target variable at first is given as the actual rental price as shown in Figure 1. To remove the skewness of the price, we applied the log transform which produced a new variable called log_price as shown in Figure 1. This transformed the price distribution into being somewhat normally distributed and symmetrical [6].

DOI: 10.54254/2754-1169/2025.BL30924

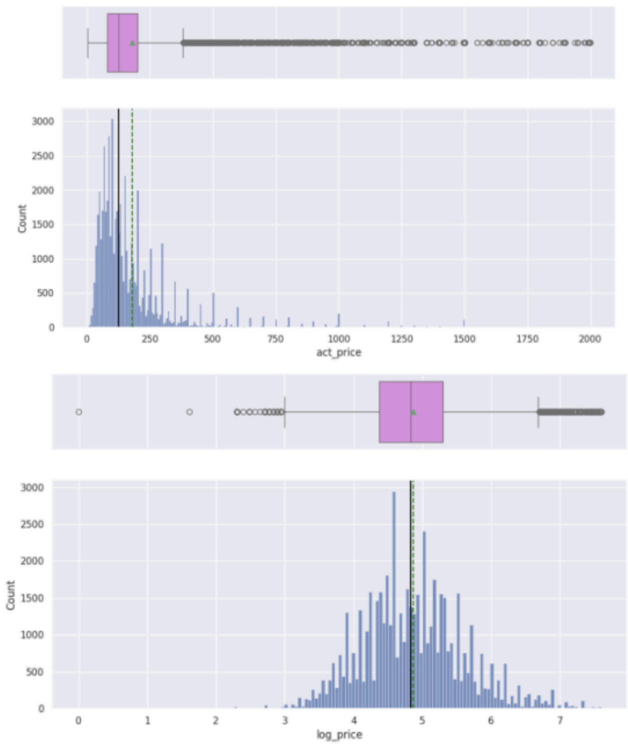

Figure 1. Box plot and histogram of act_price and log_price

Categorical feature's unique value counts were checked to find category distribution and whether encoding is necessary or not. Most were entire home/apt (33,854), then private room (18,413) and shared room (1,845). For cancellation_policy, strict type was the most popular (25,213), flexible (14,569) and moderate (14,327) types were less popular. In addition, about 74% of the listings had a cleaning_fee, and around 30% were instant_bookable..

Review_scores_rating variable is looked into and plotted as histograms and box plots as seen in Figure 2 for host performance and guest satisfaction levels. Review data can reflect the quality of the service, which will affect the price, so the variable is retained for visualization and exploratory analysis. Most of them are positive, and the lowest score in each part is 60 points. The low score could be because the guests had a bad experience.

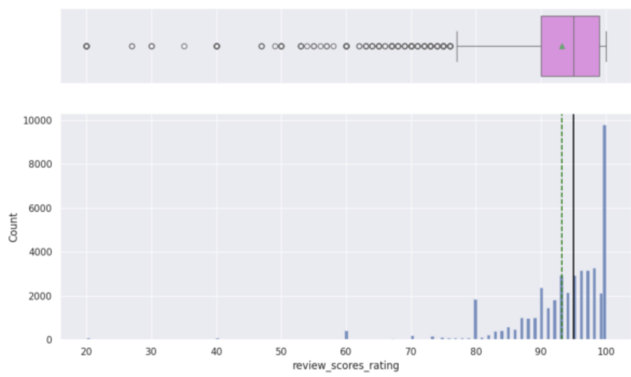

Figure 2. Box plot and histogram of review_score_rating

To check for multicollinearity of independent variables, correlation heat map was created with Python on the Airbnb data set is as given in Figure 3. , it can also be seen in the pair plot Figure 4.

DOI: 10.54254/2754-1169/2025.BL30924

Heat map visualizing the pairwise correlation of six main variables; accommodates, number of bathrooms, number of bedrooms, number of beds, review_scores_rating, and log_price. The colors range from light yellow, low correlation, to dark purple, high correlation. Results reveal that accommodates is highly correlated with bedrooms (r = 0.71) and beds (r = 0.80) as well as bedrooms is highly correlated with beds (r = 0.71). A correlation greater than 0.7 suggests that there is a fairly high correlation; thus, it could be concerning should there be several such correlations for the independent variables in the same model. To check this, the VIF test is applied to all predictors. The VIF results indicate that accommodates and beds both have VIF values over 10, which means there is multicollinearity between these two variables. High VIF values signal redundant info that can increase SE and lower stat significance of reg coeffs. Therefore, to improve model stability and interpretation, the beds variable will be removed from the regression model. The accommodates variable was kept to serve as a more direct measure of hosting capacity, and a succinct stand-in for the size of a listing's property, reducing collinearity among predictors [7,8].

Figure 3. Heat map among six key variables

Figure 4. Pair plot among six key variables

Examined data types of all variables to ensure correct formats for numerical and categorical attributes. Numerical variables such as accommodates, bathrooms, bedrooms, and act_price are stored as float values, whereas categorical variables such as room_type, cancellation_policy,

DOI: 10.54254/2754-1169/2025.BL30924

cleaning_fee, and instant_bookable are stored as objects or boolean. Correct data types were needed to continue with encoding and regression. It can be seen from Table 2 that the data types of cleaning_fee and instant_bookable have been changed from boolean to integer. Cleaning_fee and instant_bookable are specific integers that either have a value of 1 or 0, with a value of 1 representing "Yes" (i.e. a cleaning fee is charged, instant book is possible) and 0 representing "No" (i.e. a cleaning fee is not charged, instant book is not possible). This conversion makes it possible to do things like machine learning models and statistics after this. It means these features can work with the other numbers that are just simple data. We can use them right away when we make models or look at how they relate to each other [9].

Table 2. Example of variable formats after data type conversion

	room_type	accommodates	bathrooms	cancellation_policy	cleaning_fee	instant_bookable	review_scores_rating	bedrooms	beds	log_price
0	Entire home/apt	3.0	1.0	strict	1	0	100.0	1.0	1.0	5.010636
1	Entire home/apt	7.0	1.0	strict	1	1	93.0	3.0	3.0	5.129900
2	Entire home/apt	5.0	1.0	moderate	1	1	92.0	1.0	3.0	4.976735
3	Entire home/apt	4.0	1.0	flexible	1	0	96.0	2.0	2.0	6.620074
4	Entire home/apt	2.0	1.0	moderate	1	1	40.0	0.0	1.0	4.744933

From Figure 5, it can be seen that the box plot structures of various features, including accommodates, bathrooms, cleaning_fee, instant_bookable, review_scores_rating, bedrooms, beds, and log_price, are relatively regular. Although there are some outliers, they are distributed relatively reasonably within the value range of their respective features, and there are no obvious extreme outliers that deviate significantly from the overall data distribution trend.

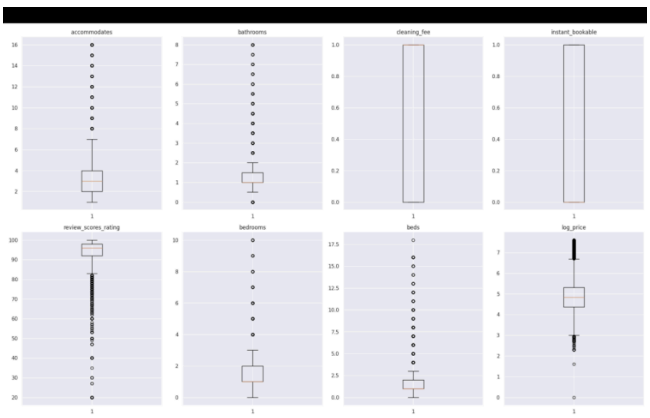

Figure 5. Box plots of six key variables

3.3. Linear regression

Linear Regression is the fundamental statistical model employed in this study for Airbnb listing price prediction, with its core objective of exploring the linear relationship between listing

DOI: 10.54254/2754-1169/2025.BL30924

characteristics and pricing (log-transformed price). The operational principle of this model is rooted in Ordinary Least Squares (OLS) estimation, which aims to determine the optimal coefficient vector by minimizing the sum of squared residuals (SSR). Mathematically, the model for the $i-th$ listing is defined as: $logy_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i$, where the dependent variable denotes the log-price of the $i-th$ listing; $x_{i1}, x_{i2}, \ldots, x_{ik}$ represent independent variables (e.g., number of beds, room_type, cleaning_fee, etc.); $\beta_0 (intercept) and \beta_1, \ldots, \beta_k$ (coefficients) are parameters to be estimated; ϵ_i is the random error term (capturing unobserved factors affecting price).

The core of OLS is to find the coefficient vector $\left(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k\right)^T$ that minimizes the SSR, as formula: $\left(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k\right)^T = \arg \min_{\beta_0, \beta_1, \ldots, \beta_k} \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})\right]^2$

This minimization ensures the "best-fitting" linear relationship—specifically, it minimizes the total residual between the model's predicted log-price $\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \ldots + \widehat{\beta}_k x_{ik}$ and the actual log-price y_i .

In matrix form, the model structures all observations into a unified framework (denoted as $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$): $\boldsymbol{Y} = [y_1, y_2, \ldots, y_n]^T$ (log-price vector for all listings); $\boldsymbol{X}$ (design matrix) is an $n \times (k+1)$ matrix containing a column of 1s (for the intercept β_0) and columns of independent variables; $\beta = [\beta_0, \beta_1, \ldots, \beta_k]^T$ (coefficient vector); $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2, \ldots, \epsilon_n]^T$ (error term vector).

After estimating β, the model's performance is evaluated using three key metrics:

• Root Mean Squared Error (RMSE): Measures the average magnitude of prediction error (sensitive to large deviations): RMSE=$\sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\hat{y}_i - y_i\right)^2}$

• Mean Absolute Error (MAE): Measures the average absolute prediction error (robust to outliers): MAE=$\frac{1}{n} \sum_{i=1}^{n} \left|\hat{y}_i - y_i\right|$

• Mean Absolute Percentage Error (MAPE): Measures the average relative prediction error (interpretable as a percentage): MAPE=$\frac{1}{n} \sum_{i=1}^{n} \left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\%$

• The coefficient of determination (R^2): It quantifies the proportion of log-price variance explained by the model: $R^2 = 1 - \frac{\mathrm{RSS_{full}}}{\mathrm{RSS_{int}}}$, where $\mathrm{RSS_{full}}$ is the SSR of the fitted model, and $\mathrm{RSS_{int}}$ is the SSR of the intercept-only model. An R^2 close to 1 indicates the model explains most of the log-price variability.

For this study's Airbnb dataset, these components collectively enable the model to capture linear patterns between listing characteristics and log-price, while the OLS minimization ensures the most accurate linear fit for prediction [10].

4. Model building and results

A multiple linear regression model was constructed to quantify the linear relationship between Airbnb listing attributes and log-transformed prices (log_price). For each independent variable, the null hypothesis $H_0 : \beta_j = 0$ was tested using p-values. A significance level of α=0.05 was adopted, with results summarized as Table 3.

DOI: 10.54254/2754-1169/2025.BL30924

Table 3. Regression results for airbnb listing attributes and log-price

Coefficient	Estimate	Std. Error	t-value	p-value
room_type	0.6834	0.006	110.278	<2.2e-16 ***
accommodates	0.0574	0.002	32.713	<2.2e-16 ***
bathrooms	0.1245	0.005	25.005	<2.2e-16 ***
cleaning_fee	-0.0748	0.006	-11.691	<2.2e-16 ***
instant_bookable	3.2448	0.033	98.929	<2.2e-16 ***
review_score_rating	0.0075	0.000	21.877	<2.2e-16 ***
bedrooms	0.1453	0.004	34.369	<2.2e-16 ***
policy_strict	-0.0089	0.007	-1.329	0.184
policy_moderate	-0.0577	0.007	-7.871	<2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.0 '*' 0.05 '.' 0.1 ' ' 1

The signs and sizes of the coefficients show the direction and size of the impact of the independent variables on log_price. As we have log transformed our price, these coefficients are interpreted as the average change in log_price for one unit increase in the independent variables. Among which, the coefficient of "instant_bookable" (3.2448) is the largest of all the variables, meaning that instantly bookable listings have the greatest effect on increasing the log_price by about 3.24, having the most significant positive impact on the price. 0.6834 is the coefficient for room_type, which means that room type enhancements raise the log_price by roughly 0.68 on average. Variable “bedrooms” and “bathrooms” has small coefficients, they have weak positive impact to price. Among the variables that have negative effect are “cleaning_fee” (-0.0748) which means that every one unit increase in cleaning fees causes a reduction of log_price by an average of about 0.07 and “policy_moderate” (-0.0577) which means that moderately strict policy causes a reduction of log_price by an average of about 0.06 [11]. Concerning the significance tests of each variable, apart from “policy_strict”, the p-values of all other variables are very low, which is less than 0.05 (less than 2.2e - 16). This means that there is a linear relationship between these variables and log_price. But the p value for policy_strict is 0.184 which is greater than 0.05. Therefore, we fail to reject the null hypothesis (H_0: β=0). This means that policy_strict does not have a statistically significant effect on log_price, and it is weak in its capacity to explain the variability of log_price and contributes minimally to predicting log_prices [12,13].

The model's R^2 value is 0.486. Its core form involves taking the logarithm of the dependent variable $\boldsymbol{Y}$ ($ln(\boldsymbol{Y})$) and performing linear regression with the independent variable X ($\ln(Y) = \beta_0 + \beta_1 X + u$). Here, the model's "dependent variable" is actually $ln(\boldsymbol{Y})$, and R^2 is calculated for this logarithmically transformed dependent variable. An R^2 value of 0.486 indicates that the model explains 48.6% of the variation in log_price. This suggests that these property attributes account for nearly half of the price fluctuations, with the remaining variation explained by other factors not included in the model, such as geographic location, interior design style, etc. [14].

5. Discussion and conclusion

In this paper, I use Multiple linear regression model to see the linear relations between Airbnb listing variables and logarhythm of price(log_price), so as to find the important factors which are related to listings’ price, so as to give references to Airbnb hosts for setting up its own prices. The model's R^2

DOI: 10.54254/2754-1169/2025.BL30924

value is 0.486, meaning that the selected listing attributes explain 48.6% of the variation in log_price. This result suggests that a great part of the variance is explained and the pricing is jointly driven by measurable listing attributes and unobservable factors [14]. The variable “instant_bookable” is most influential from the point of view of its coefficients among all other variables. The coefficient of “instant_bookable” has the highest value i.e. 3.2448 and its p-value is much lower than 0.05 ($p<2.2e-16$ ***). Therefore, we can conclude that instant_bookable has a significant positive impact on log_price. This shows that the instant booking service greatly improves the tenant's booking convenience, which satisfies the consumers' need for fast and flexible accommodation. Thus, hosts who enable the instant booking function are able to acquire a pricing premium and service convenience is an important factor for sharing accommodations pricing, as demonstrated by the research. The second most influential factor was “room_type” with a coefficient of 0.6834 and $p < 2.2e - 16$ *** which indicates that a higher quality, or a more premium room type, greatly increases the price. This is due to the fact that better room types will be able to offer a more private and comfortable space for living, which is an important factor for tenants to choose a premium accommodation. variables like “bedrooms” and “bathrooms” also have quite a large impact on the log of the price, however their coefficients are relatively small at 0.1453 and 0.1245 respectively. It indicates that the number of bedrooms and bathrooms are a basic guarantee of accommodation quality, but their effect on price increase is less than the basic attributes of instant booking and room type grade [11-13]. In the variables that have negative impact “cleaning_fee” (coefficient = -0.0748, $p < 2.2e-16$ ***) and “policy_moderate” (coefficient = -0.0577, $p < 2.2e-16$ ***) passed the significance test. Cleaning_fee may have a -ve correlation with log_price as,when tenants find an increase in the cost of clean,they are more price sensitive and as such hosts lower the base price moderately to remain competitive. Negative effect of moderate cancellation policy on price shows that tenants are more flexible with their booking. Moderately strict policies increase the risk of a tenant changing their reservation, and hosts have to lower their price to compensate for this risk. On the other hand, “policy_strict” has a p - value of 0.184 that is greater than the significance level of 0.05, therefore, it is not statistically significant. It might be due to strict cancellation policies which have a double-edged effect on tenants, by increasing the risk of tenants’ reservations, but reducing the hosts’ operational uncertainty [11,13].

And this study also has certain limits. Firstly, the amount of data is insufficient, the data only covers New York, but there is no data on other regions, which cannot fully cover the listing diversity among different regions, markets, and operating models, thus leading to biases in the research conclusions. Secondly, there is a rather rough processing of the data: for the missing values, the median method was used for filling, but no analysis of the cause of the missing value occurred, and a more complex way of filling the missing value may be selected; For the category variables, the way of encoding is simple and cannot completely reflect the relationship between the categories and the interaction between them. Third, model construction is too simple. Linear regression can only reflect the linear relationship between variables, but the actual impact of listing features on pricing may have nonlinear correlations and complex interrelations among factors that the current model cannot fully capture, thus affecting the overall explanation and prediction accuracy. Future research can overcome these shortcomings with the targeted optimization. Data, expanding the amount of data to include cross-regional, multi-temporal listing data, and increasing the sample size to tens of thousands or even hundreds of thousands can make the research more representative and robust. In data processing, using better methods like multiple imputations for missing data, target encoding/ embeddings for categories, and domain-knowledge driven outlier handling will make the data better. When selecting a model, incorporating newer types of algorithms like random forests, gradient

DOI: 10.54254/2754-1169/2025.BL30924

boosting machine can catch nonlinear relations and interactions among variables. At the same time, future research can further increase the number of dimensions of the variables used in the study, such as spatial location indicator, host indicators, and social trust indicator, to build a more complete mechanism model on pricing impact mechanism, so as to provide more targeted and deepened insights on sustainable development of the Airbnb market.

References

[1] Samwel, M. (2022). What Factors drives the Airbnb Listing's Prices? International Business & Economics Studies, 4(1), p26. https: //doi.org/10.22158/ibes.v4n1p26
[2] Cai, Y., Zhou, Y., Jenny) MA, J., & Scott, N. (2019). Price Determinants of Airbnb Listings: Evidence from Hong Kong. Tourism Analysis, 24(2), 227–242. https: //doi.org/10.3727/108354219x15525055915554
[3] Claanen. (2021, May 11). Neighborhood Change and Residential Instability in Oakland. Federal Reserve Bank of San Francisco; San Francisco Fed. https: //www.frbsf.org/research-and-insights/publications/community-development-working-papers/2021/05/neighborhood-change-and-residential-instability-in-oakland/
[4] Quattrone, G., Greatorex, A., Quercia, D., Capra, L., & Musolesi, M. (2018). Analyzing and predicting the spatial penetration of Airbnb in U.S. cities. EPJ Data Science, 7(1). https: //doi.org/10.1140/epjds/s13688-018-0156-6
[5] Torres, F. (2024, December 30). Interpretability and Transparency of Machine Learning Models. Medium. https: //medium.com/@fertorresfs/interpretability-and-transparency-of-machine-learning-models-b1f169ae582d
[6] Codecademy. (2017). Log Transformation in Linear Regression: When and How to Use It. Codecademy. https: //www.codecademy.com/article/log-transformation-in-linear-regression-when-and-how-to-use-it
[7] Frost, J. (2017, April 2). Multicollinearity in Regression Analysis: Problems, Detection, and Solutions. Statistics by Jim. https: //statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/#google_vignette
[8] Pennsylvania State University. (2018). 10.7 - Detecting Multicollinearity Using Variance Inflation Factors | STAT 462. Online.stat.psu.edu. https: //online.stat.psu.edu/stat462/node/180/
[9] Jaiswal, S. (2022, December). Data Type Conversion in Python Tutorial. Www.datacamp.com. https: //www.datacamp.com/tutorial/python-data-type-conversion
[10] GeeksforGeeks. (2018, September 13). Linear Regression in Machine learning. GeeksforGeeks. https: //www.geeksforgeeks.org/machine-learning/ml-linear-regression/
[11] GeeksforGeeks. (2024b, May 16). Regression Coefficients. GeeksforGeeks. https: //www.geeksforgeeks.org/maths/regression-coefficients/
[12] S.3.2 Hypothesis Testing (P-Value Approach) | STAT ONLINE. (n.d.-b). PennState: Statistics Online Courses. https: //online.stat.psu.edu/statprogram/reviews/statistical-concepts/hypothesis-testing/p-value-approach
[13] Frost, J. (n.d.). How to Interpret P-values and Coefficients in Regression Analysis. Statistics by Jim. https: //statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/
[14] Kosourova, E. (2025, May 14). R-Squared Explained: How Well Does Your Regression Model Fit? Datacamp.com; DataCamp. https: //www.datacamp.com/tutorial/r-squared