MERCER UNIVERSITY

Stetson-Hatcher School of Business

Airbnb Analytics: Predicting Rental Prices

Submitted by

PRABHU SHANKAR GANDASI VIRUPAKSHA	11062444
JOHN POLE MADHU	11062542
RAJIVINI TIRUVEEDHULA	11056828

1. INTRODUCTION

The tourism industry has grown exponentially in the past few years driven by factors such as globalization, rising disposable incomes, advancements in technology, and evolving consumer preferences. This growth has brought significant economic, cultural, and social benefits while also presenting challenges that require strategic management. This is where Airbnb recognised the gap in the industry and tackled it, specifically in the context of accommodations.

Airbnb, founded in 2008, is an online marketplace that connects people looking to rent their homes with travellers seeking accommodations. Airbnb offers unique stays, including entire homes, apartments, shared spaces, and unconventional properties like treehouses, boats, and yurts. It caters to travellers seeking more personalized and authentic experiences compared to traditional hotels. In the broader travel industry, Airbnb represents the shift towards more personalized, flexible, and technology-driven travel solutions, significantly influencing how people explore the world.

Airbnb listings are priced based on factors such as size, location, amenities, features, and so on.

Business Question: How can Airbnb hosts and potential renters leverage data to accurately predict rental prices, optimize revenue, and enhance decision-making in a competitive marketplace?

We aim to understand how different factors about various listings determine the prices of the properties and how potential hosts can leverage this data to predict an optimal price for their listings so that they can price their property at an appropriate price to stand out in the market while maximizing their revenue.

How does this help hosts?

Determining optimal pricing for their property based on factors like location, amenities, number of people accommodated etc. When setting the price for a rental property, several factors come into play. Ensuring the pricing strategy aligns with market trends, guest expectations, and operational goals is critical for maximising revenue while attracting bookings.

It is also important for hosts to understand how competitively price their listings to position their listing attractively.

2. LITERATURE REVIEW

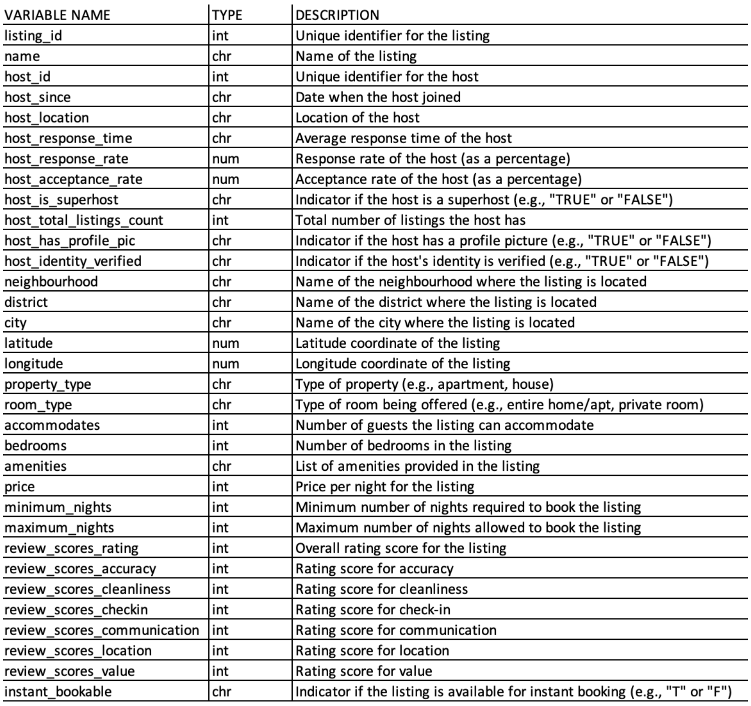
Study 1: This study aims to assess the impact of various factors impacting prices. It applies OLS regression to assess the relationship between price and other factors. It aims to create instruments that will help to predict and plan prices of the Airbnbs keeping tourism stakeholders and policymakers in mind. (What Factors Drive the Airbnb Listing's Prices? By Samwel Meigeka)

Study 2: This paper tries to understand and investigate the effect of price discrimination on Airbnb listing revenue. It tries to study the magnitude of these effects and shed light on the listing attributes that cause a price surge, specifically impacting host revenue. (Dynamic pricing and revenues of Airbnb listings by Veronica Leoni)

3. DATA

3.1 DATA COLLECTION

The original dataset has 279712 observations and 31 variables. A description of the variables is given in the table below:

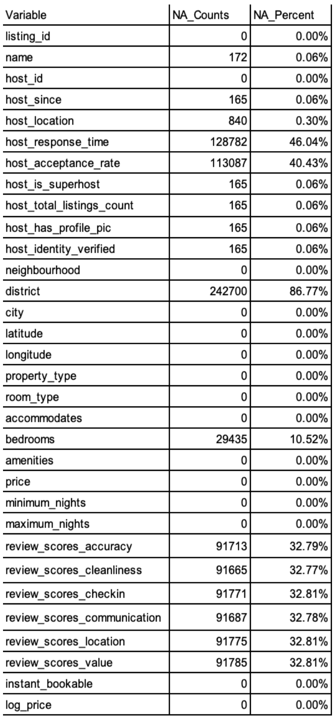


VARIABLE NAME	TYPE	DESCRIPTION
listing_id	int	Unique identifier for the listing
name	chr	Name of the listing
host_id	int	Unique identifier for the host
host_since	chr	Date when the host joined
host_location	chr	Location of the host
host_response_time	chr	Average response time of the host
host_response_rate	num	Response rate of the host (as a percentage)
host_acceptance_rate	num	Acceptance rate of the host (as a percentage)
host_is_superhost	chr	Indicator if the host is a superhost (e.g., "TRUE" or "FALSE")
host_total_listings_count	int	Total number of listings the host has
host_has_profile_pic	chr	Indicator if the host has a profile picture (e.g., "TRUE" or "FALSE")
host_identity_verified	chr	Indicator if the host's identity is verified (e.g., "TRUE" or "FALSE")
neighbourhood	chr	Name of the neighbourhood where the listing is located
district	chr	Name of the district where the listing is located
city	chr	Name of the city where the listing is located
latitude	num	Latitude coordinate of the listing
longitude	num	Longitude coordinate of the listing
property_type	chr	Type of property (e.g., apartment, house)
room_type	chr	Type of room being offered (e.g., entire home/apt, private room)
accommodates	int	Number of guests the listing can accommodate
bedrooms	int	Number of bedrooms in the listing
amenities	chr	List of amenities provided in the listing
price	int	Price per night for the listing
minimum_nights	int	Minimum number of nights required to book the listing
maximum_nights	int	Maximum number of nights allowed to book the listing
review_scores_rating	int	Overall rating score for the listing
review_scores_accuracy	int	Rating score for accuracy
review_scores_cleanliness	int	Rating score for cleanliness
review_scores_checkin	int	Rating score for check-in
review_scores_communication	int	Rating score for communication
review_scores_location	int	Rating score for location
review_scores_value	int	Rating score for value
instant_bookable	chr	Indicator if the listing is available for instant booking (e.g., "T" or "F")

3.2 DATA PREPROCESSING

Preprocessing was done by checking for the outliers and checking for NA values.

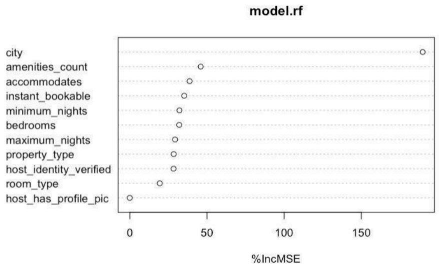
The NA values % is given in the below table:



Variable	NA_Counts	NA_Percent
listing_id	0	0.00%
name	172	0.06%
host_id	0	0.00%
host_since	165	0.06%
host_location	840	0.30%
host_response_time	128782	46.04%
host_acceptance_rate	113087	40.43%
host_is_superhost	165	0.06%
host_total_listings_count	165	0.06%
host_has_profile_pic	165	0.06%
host_identity_verified	165	0.06%
neighbourhood	0	0.00%
district	242700	86.77%
city	0	0.00%
latitude	0	0.00%
longitude	0	0.00%
property_type	0	0.00%
room_type	0	0.00%
accommodates	0	0.00%
bedrooms	29435	10.52%
amenities	0	0.00%
price	0	0.00%
minimum_nights	0	0.00%
maximum_nights	0	0.00%
review_scores_accuracy	91713	32.79%
review_scores_cleanliness	91665	32.77%
review_scores_checkin	91771	32.81%
review_scores_communication	91687	32.78%
review_scores_location	91775	32.81%
review_scores_value	91785	32.81%
instant_bookable	0	0.00%
log_price	0	0.00%

The high number of NA values in a few of the columns stated the data was uneven and the below steps were taken to clean them.

We ran regression models to see which are the significant variables and then we proceeded with the imputation. Below are the significant variables



3.3 DATA DIMENSION REDUCTION

We were able to reduce the variables by domain knowledge and significant variables we obtained using the random forest and the linear regression.

3.4 PREDICTION OVERVIEW

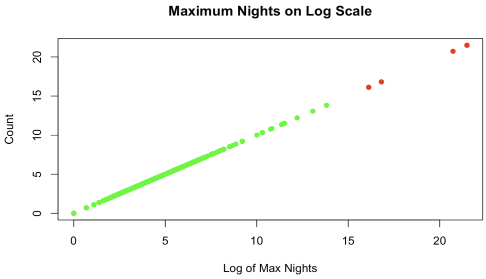
In this project, we aim to predict the price of accommodations based on a range of key features, leveraging a log-linear regression model to better capture the relationships between variables and price. The analysis includes room attributes, host characteristics, and location-specific factors to determine their impact on price while ensuring statistical rigor and interpretability of results. Below is a detailed breakdown of the significant predictors and their effects.

3.5 DATA CLEANING

Observation of the maximum nights from the below chart shows there are massive outliers.

The interquartile method was used to clean all the variables, with the imputation of mean to the number of bedrooms as it is a numerical variable.

For better understanding, we made a chart with log prices.



3.6 DATA PARTITIONING

We used a partitioning of 80-20 of the original dataset. 80% was given to training, and 20% was to testing.

HEDONIC PRICING MODEL:

The hedonic pricing model applies linear regression to estimate the value of an item based on its characteristics. It is commonly used in real estate to determine the factors that contribute to a property's value. The model analyzes various attributes of a property, such as size and location, to assess how each factor influences the price. While there are more complex methods available, this study focuses on the straightforward linear regression approach.

4. METHODOLOGY

4.1 VARIABLE SELECTION

The initial dataset had 33 variables which is too large for analysis. Therefore, we decided to create a subset of variables that will help us make a meaningful prediction of the price and which may directly influence the pricing of each listing. Variables were selected based on domain knowledge and by performing a random forest to see which variables are the most important and a correlation matrix was use to see the relationship between the predictors.

After this process, we concluded our model to have 11 variables which are: room_type Accommodates, bedrooms, minimum_nights, minimum_nights, maximum_nights,
City.

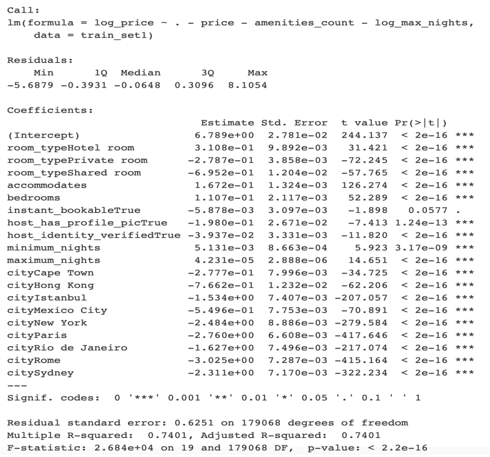
4.2 MODELS

4.2.1 Baseline Model

Before running our models, it is important to create a baseline model, which will act as a standard for checking all other models against. Our baseline model was created by simply taking the averages and predicting what our threshold should be for all the models we will be running.

4.2.2 Linear regression

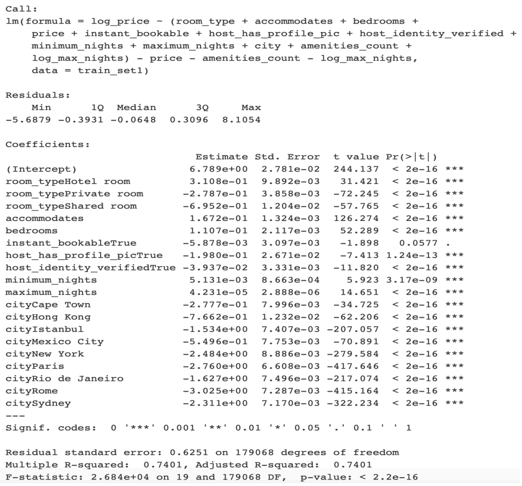
After running the baseline model, we need to assess how our variables would perform to predict the prices. The initial linear model was run using log_price as the dependent variable and all the other variables except price(as log price was used), log_max_nights and amenities_count as the predictor variables. The results are shown below:



```
Call:
lm(formula = log_price ~ . - price - amenities_count - log_max_nights,
    data = train_set1)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.6879 -0.3931 -0.0648  0.3096  8.1054 

Coefficients:
                            Estimate Std. Error  t value Pr(>|t|)    
(Intercept)                6.789e+00  2.781e-02  244.137  < 2e-16 ***
room_typeHotel room        3.108e-01  9.892e-03   31.421  < 2e-16 ***
room_typePrivate room     -2.787e-01  3.858e-03  -72.245  < 2e-16 ***
room_typeShared room      -6.952e-01  1.204e-02  -57.765  < 2e-16 ***
accommodates               1.672e-01  1.324e-03  126.274  < 2e-16 ***
bedrooms                   1.107e-01  2.117e-03   52.289  < 2e-16 ***
instant_bookableTrue      -5.878e-03  3.097e-03   -1.898   0.0577 .  
host_has_profile_picTrue  -1.980e-01  2.671e-02   -7.413 1.24e-13 ***
host_identity_verifiedTrue -3.937e-02  3.331e-03  -11.820  < 2e-16 ***
minimum_nights             5.131e-03  8.663e-04    5.923 3.17e-09 ***
maximum_nights             4.231e-05  2.888e-06   14.651  < 2e-16 ***
cityCape Town             -2.777e-01  7.996e-03  -34.725  < 2e-16 ***
cityHong Kong             -7.662e-01  1.232e-02  -62.206  < 2e-16 ***
cityIstanbul              -1.534e+00  7.407e-03 -207.057  < 2e-16 ***
cityMexico City           -5.496e-01  7.753e-03  -70.891  < 2e-16 ***
cityNew York              -2.484e+00  8.886e-03 -279.584  < 2e-16 ***
cityParis                 -2.760e+00  6.608e-03 -417.646  < 2e-16 ***
cityRio de Janeiro        -1.627e+00  7.496e-03 -217.074  < 2e-16 ***
cityRome                  -3.025e+00  7.287e-03 -415.164  < 2e-16 ***
citySydney                -2.311e+00  7.170e-03 -322.234  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6251 on 179068 degrees of freedom
Multiple R-squared:  0.7401, Adjusted R-squared:  0.7401 
F-statistic: 2.684e+04 on 19 and 179068 DF,  p-value: < 2.2e-16
```

4.2.3 Stepwise Model

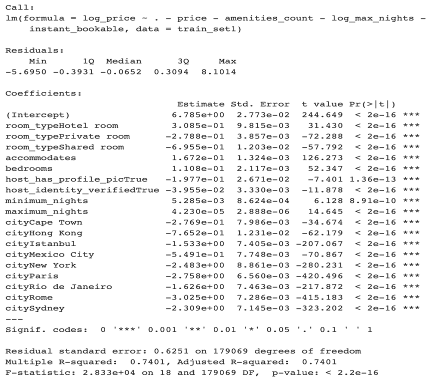
After running the initial linear model, we ran a stepwise model to see which variables are significant and which variables are significant. After running the stepwise model, we can see that the variable instant bookable was eliminated. The results are shown below:



```
Call:
lm(formula = log_price ~ (room_type + accommodates + bedrooms + 
    price + instant_bookable + host_has_profile_pic + host_identity_verified + 
    minimum_nights + maximum_nights + city + amenities_count + 
    log_max_nights) - price - amenities_count - log_max_nights, 
    data = train_set1)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.6879 -0.3931 -0.0648  0.3096  8.1054 

Coefficients:
                            Estimate Std. Error  t value Pr(>|t|)    
(Intercept)                6.789e+00  2.781e-02  244.137  < 2e-16 ***
room_typeHotel room        3.108e-01  9.892e-03   31.421  < 2e-16 ***
room_typePrivate room     -2.787e-01  3.858e-03  -72.245  < 2e-16 ***
room_typeShared room      -6.952e-01  1.204e-02  -57.765  < 2e-16 ***
accommodates               1.672e-01  1.324e-03  126.274  < 2e-16 ***
bedrooms                   1.107e-01  2.117e-03   52.289  < 2e-16 ***
instant_bookableTrue      -5.878e-03  3.097e-03   -1.898   0.0577 .  
host_has_profile_picTrue  -1.980e-01  2.671e-02   -7.413 1.24e-13 ***
host_identity_verifiedTrue -3.937e-02  3.331e-03  -11.820  < 2e-16 ***
minimum_nights             5.131e-03  8.663e-04    5.923 3.17e-09 ***
maximum_nights             4.231e-05  2.888e-06   14.651  < 2e-16 ***
cityCape Town             -2.777e-01  7.996e-03  -34.725  < 2e-16 ***
cityHong Kong             -7.662e-01  1.232e-02  -62.206  < 2e-16 ***
cityIstanbul              -1.534e+00  7.407e-03 -207.057  < 2e-16 ***
cityMexico City           -5.496e-01  7.753e-03  -70.891  < 2e-16 ***
cityNew York              -2.484e+00  8.886e-03 -279.584  < 2e-16 ***
cityParis                 -2.760e+00  6.608e-03 -417.646  < 2e-16 ***
cityRio de Janeiro        -1.627e+00  7.496e-03 -217.074  < 2e-16 ***
cityRome                  -3.025e+00  7.287e-03 -415.164  < 2e-16 ***
citySydney                -2.311e+00  7.170e-03 -322.234  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6251 on 179068 degrees of freedom
Multiple R-squared:  0.7401, Adjusted R-squared:  0.7401 
F-statistic: 2.684e+04 on 19 and 179068 DF,  p-value: < 2.2e-16
```

4.2.4 Linear Regression 2

After running the stepwise model, we ran another linear regression, this time by removing instant bookable so that the accuracy could improve, which improved significantly. The results are shown below:

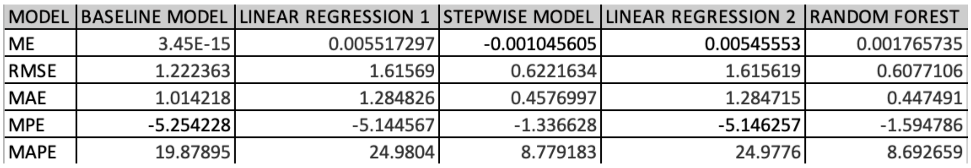


```
Call:
lm(formula = log_price ~ . - price - amenities_count - log_max_nights -
    instant_bookable, data = train_set1)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6950 -0.3931 -0.0652  0.3094  8.1014

Coefficients:
                            Estimate Std. Error  t value Pr(>|t|)
(Intercept)                6.785e+00  2.773e-02  244.649  < 2e-16 ***
room_typeHotel room        3.085e-01  9.815e-03   31.430  < 2e-16 ***
room_typePrivate room     -2.788e-01  3.857e-03  -72.288  < 2e-16 ***
room_typeShared room      -6.955e-01  1.203e-02  -57.792  < 2e-16 ***
accommodates               1.672e-01  1.324e-03  126.273  < 2e-16 ***
bedrooms                   1.108e-01  2.117e-03   52.347  < 2e-16 ***
host_has_profile_picTrue  -1.977e-01  2.671e-02   -7.401 1.36e-13 ***
host_identity_verifiedTrue -3.955e-02  3.330e-03  -11.878  < 2e-16 ***
minimum_nights             5.285e-03  8.624e-04    6.128 8.91e-10 ***
maximum_nights             4.230e-05  2.888e-06   14.645  < 2e-16 ***
cityCape Town             -2.769e-01  7.986e-03  -34.674  < 2e-16 ***
cityHong Kong             -7.652e-01  1.231e-02  -62.179  < 2e-16 ***
cityIstanbul              -1.533e+00  7.405e-03 -207.067  < 2e-16 ***
cityMexico City           -5.491e-01  7.748e-03  -70.867  < 2e-16 ***
cityNew York              -2.483e+00  8.861e-03 -280.231  < 2e-16 ***
cityParis                 -2.758e+00  6.560e-03 -420.496  < 2e-16 ***
cityRio de Janeiro        -1.626e+00  7.463e-03 -217.872  < 2e-16 ***
cityRome                  -3.025e+00  7.286e-03 -415.183  < 2e-16 ***
citySydney                -2.309e+00  7.145e-03 -323.202  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6251 on 179069 degrees of freedom
Multiple R-squared:  0.7401, Adjusted R-squared:  0.7401
F-statistic: 2.833e+04 on 18 and 179069 DF,  p-value: < 2.2e-16
```

4.2.5 Random Forest

The final model that we ran was a Random forest on the selected variables to improve the flexibility, robustness, and performance.

5. RESULTS



MODEL	BASELINE MODEL	LINEAR REGRESSION 1	STEPWISE MODEL	LINEAR REGRESSION 2	RANDOM FOREST
ME	3.45E-15	0.005517297	-0.001045605	0.00545553	0.001765735
RMSE	1.222363	1.61569	0.6221634	1.615619	0.6077106
MAE	1.014218	1.284826	0.4576997	1.284715	0.447491
MPE	-5.254228	-5.144567	-1.336628	-5.146257	-1.594786
MAPE	19.87895	24.9804	8.779183	24.9776	8.692659

Baseline Model: The baseline model provides a starting point for comparison. While its ME is close to zero, indicating no bias, its RMSE and MAE values are relatively high (1.222 and 1.014, respectively), suggesting poor predictive accuracy. The MAPE of 19.88% shows the error in percentage terms.

Linear Regression 1: The linear regression model shows comparable performance, with MPE near -5, indicating a slight underestimation in predictions. However, their RMSE (1.616) and MAE (1.285) values are not as good as the baseline model, showing that there is still scope for improvement.

Stepwise Model: The stepwise regression significantly outperforms the baseline model in terms of error reduction, with the lowest RMSE (0.622), MAE (0.458), and MAPE (8.78%). This indicates a well-tuned regression approach that captures the underlying data patterns better than the others.

Linear Regression 2: Linear Regression 2 has significantly improved after eliminating the variables. However, its error metrics (RMSE, MAE, MAPE) are not that great in comparison to the baseline model, The ME and MPE values suggest minimal bias.

Random Forest: The Random Forest model delivers the best overall performance. With the lowest RMSE (0.607) and MAE (0.447), it surpasses even the stepwise model slightly in accuracy. The MAPE (8.69%) also confirms its robustness and precision.

6. CONCLUSION AND RECOMMENDATIONS

6.1 CONCLUSION: Given the low RMSE of 0.6077, Random Forest proves to be a reliable model for predicting the price of new Airbnb listings. It effectively captures complex relationships between variables such as room type, accommodation capacity, bedrooms, and minimum nights, making it a robust tool for pricing recommendations. Hosts can confidently use this model to set competitive and accurate prices for their listings.

6.2 RECOMMENDATIONS:

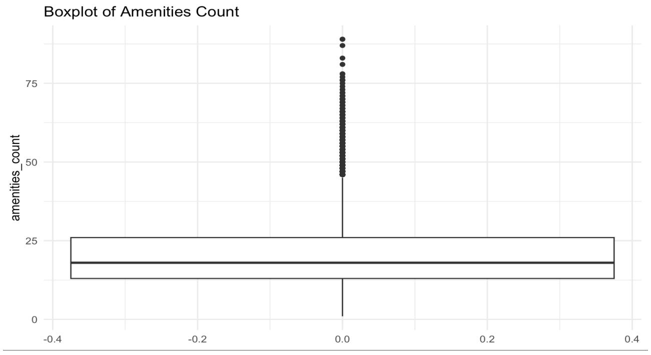
Based on the data, here are recommendations for hosts to optimize their Airbnb listings:

1. **Focus on Room Type**: Listings categorized as "Hotel room" have the highest impact on price (36.10%). If possible, consider aligning your offering with features typical of hotel rooms to justify higher pricing.
2. **Maximize Accommodation Capacity**: The number of people a property can accommodate contributes significantly (18.10%) to pricing. Ensure your listing highlights its capacity and provides adequate amenities to cater to this number.
3. **Optimize Bedroom Count**: Bedroom count accounts for 11.80% of price variation. If feasible, adding an extra bedroom or converting a space into a functional sleeping area could increase your revenue potential.
4. **Reassess Minimum Night Requirements**: Minimum nights have a minimal impact on price (0.56%). Consider lowering the minimum night stay to attract more bookings without significantly impacting revenue.

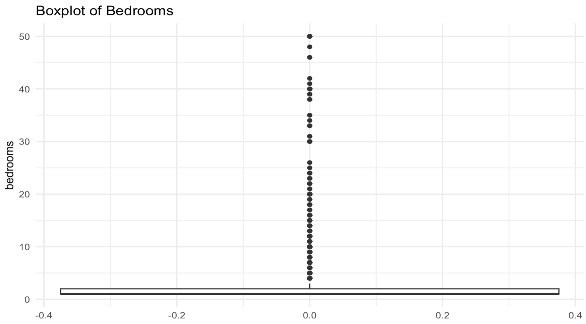
Appendix:

1. More information about EDA

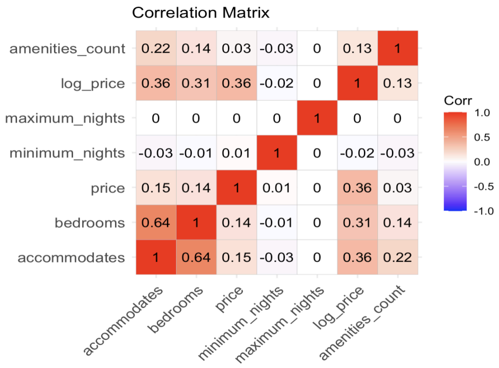
The boxplot shows the distribution of the number of amenities across listings, with the majority of listings clustered below 25 amenities. Outliers above this range indicate a few properties offering significantly more amenities.



Bedrooms: A boxplot was created to visualize the distribution of the number of bedrooms across different listings. This representation highlights the spread, central tendency, and potential outliers in the number of bedrooms, offering insights into typical property configurations and variations.

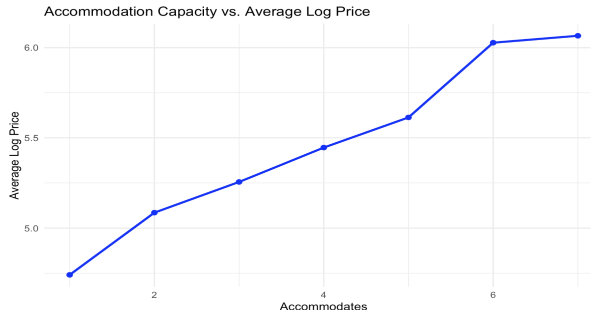


A correlation matrix was generated to identify the relationships between variables, highlighting which variables are strongly correlated. This analysis provides insights into potential dependencies and helps in determining which factors might influence each other, guiding further analysis and model development.

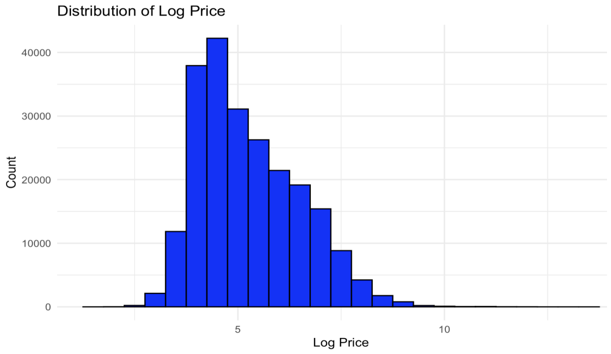


	accommodates	bedrooms	price	minimum_nights	maximum_nights	log_price	amenities_count
amenities_count	0.22	0.14	0.03	-0.03	0	0.13	1
log_price	0.36	0.31	0.36	-0.02	0	1	0.13
maximum_nights	0	0	0	0	1	0	0
minimum_nights	-0.03	-0.01	0.01	1	0	-0.02	-0.03
price	0.15	0.14	1	0.01	0	0.36	0.03
bedrooms	0.64	1	0.14	-0.01	0	0.31	0.14
accommodates	1	0.64	0.15	-0.03	0	0.36	0.22

The graph below illustrates the relationship between the number of accommodations and the average price. This analysis aims to identify any potential correlation or trend that might exist between these two variables, providing insights into how accommodation availability impacts pricing strategies.



The histogram shows the distribution of log-transformed prices for Airbnb listings, with the majority of listings centred around a log price of 5–6. The right tail indicates a smaller number of higher-priced listings, highlighting a skewed distribution before transformation.



References

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

Leoni, Veronica & Nilsson, William. (2021). Dynamic pricing and revenues of Airbnb listings: Estimating heterogeneous causal effects. International Journal of Hospitality Management. 95. 10.1016/j.ijhm.2021.102914.

OpenDataSoft (2023). Airbnb Listings. Available at: https://data.opendatasoft.com/explore/dataset/airbnb- listings%40public/table/?disjunctive.host-verificationsdisjunctive.amenities&disjunctive.features

Opendatasoft (2023) Customers - Opendatasoft. Available at: https://www.opendatasoft.com/en/customers/