IS630 – Statistical Thinking in Data Science

Home Away from Home: An Explanatory Analysis of the Key Drivers of Airbnb Occupancy

6th November 2025

Names: Tanisha Kaur, Priya Latha, Bao Sihan, Bai Haotian, Hor Wen Xuan, Serene Cheng

Content Page

SMU
SINGAPORE MANAGEMENT UNIVERSITY

School of
Information Systems

1. Introduction

The rise of Airbnb has revolutionised the hospitality industry across the world, transforming how tourists find accommodations and allowing property owners to generate income from their spaces. As of today, Airbnb boasts over 8 million listings in over 200 countries (Kumar, 2025), creating a highly competitive market.

This project investigates the key drivers of success behind Airbnb's listings through analysis of factors such as location, property features, host behaviour, guest experience metrics, pricing and availability settings from a dataset of over 10,000 Airbnb listings. Success is defined as Airbnb listings with the highest occupancy rates. Our project aims to understand what factors result in high occupancy rates and generate effective recommendations that strengthen Airbnb's revenue strategies and allow hosts to optimise their listings.

2. Background & Objective of Study

2.1 Problem Statement

According to the United Nations World Tourism Organization (2025), international tourist arrivals increased by 5% in the first six months of 2025 compared to the same period in 2024. Due to the popularity of Airbnbs, more travelers have been shifting away from hotels to Airbnb accommodations (Chen & Schuckert, 2016). However, despite the rising demand, Airbnb hosts are still unable to identify the listing attributes that are associated with higher occupancy rates. This gap makes it challenging for hosts to identify the variables that are most closely associated with occupancy.

2.2 Novelty

Based on existing research, there has been more focus on guest satisfaction rather than determinants that understand the variation of occupancy among listings. In order to address this gap, our group has decided to approach this project from a different angle as discussed in the following paragraphs.

2.3 Objective

The objective of our analysis is to identify which variables in our dataset are associated with higher occupancy rates. It is important to note that occupancy was derived from the number of reviews. The insights generated from the project aims to provide Airbnb with a clearer perspective to strengthen revenue strategies as well as provide information to hosts to optimise their listings.

2.4 Motivation

Our group is fascinated by the digital platform business models, and the hospitality industry is close to heart. We are intrigued by the opportunity to take on the perspective of Airbnb and understand the variables that are most strongly associated with the occupancy of a listing. We believe our findings will provide Airbnb with the knowledge on the key factors that drive occupancy. This project provided us with the flexibility to merge our passion and academic desire to understand the variables that shape the attractiveness of a listing.

2.5 Analytical Question

Our report aims to address the following:

What variables are associated with high Airbnb occupancy rates?

Answering this question is important as it helps identify which factors are associated with high occupancy rates on Airbnb. This information can guide hosts in optimising their listings for higher occupancy and ratings, and facilitate Airbnb in improving host education.

3. Data Sources

3.1 Dataset and Variables

This project deploys annual data from Airbnbs in Sydney, Australia. This was obtained from Inside Airbnb, a website that sources publicly available information from Airbnb. The dataset captures variables that could determine a listing's overall attractiveness, as many of the variables reflect the information that users have when selecting a listing.

The dataset is large, with 79 columns and 18,187 rows of data. As such, we have filtered for the most relevant variables, largely categorised into: host information, property attractiveness, location and performance metrics. The final list of variables is shown below:

Variable	Description	Data type	Proposed use
Estimated Occupancy Rate	Estimated proportion of days the listing was booked over the past year	Numerical continuous	Dependent variable
Host Response Rate	Percentage of guest messages that host replies to within 24 hours	Numerical continuous	To assess whether responsiveness of host is associated with occupancy rate
Host Response Time	How quickly host responds to guest inquiries or booking requests	Categorical	

Host Acceptance Rate	Percentage of booking requests accepted by host	Numerical continuous	Assess whether host acceptance rate has an association with occupancy rate
Instant Bookable	Whether listing can be booked instantly without host approval	Boolean	To evaluate whether instant bookable listing has an association with occupancy rate
Host Identity Verified	Whether host's identity is verified	Boolean	To assess whether credibility of the host is associated with occupancy rate
Host is Superhost	Whether host has Superhost status	Boolean	
License	Whether listing has verified license	Boolean	
Overall Review Score	The overall review rating of listing	Numerical continuous	
Room Type	Type of space offered (e.g., Entire home, Private room)	Categorical	To assess how listing characteristic is associated with occupancy rate
Number of Bathrooms	Number of bathroom(s) in listing	Numerical continuous	
Number of Bedrooms	Number of bedroom(s) in listing	Numerical discrete	
Minimum Nights Required	Minimum number of nights required per booking	Numerical discrete	
Maximum Nights Allowed	Maximum number of nights allowed per booking		
Accomodates	Maximum capacity of listing	Numerical discrete	
Count of amenities	List of number of amenities offered by the listing	Numerical discrete	Identifying whether the number of amenities in a listing has an association with occupancy rate
Price (per night)	Daily price of the listing in AUD	Numerical continuous	Assessing the association between

			price set by host and occupancy rate
First review received	Date of the first review received for the listing	Date	Used to exclude new listings with insufficient booking history
Last review received	Date of the most recent review received for the listing	Date	

3.2 Data Quality and Preparation

1. Filtering listings with insufficient booking history

Some listings may have been created less than a year ago, therefore listings with a first review received less than one year ago were excluded to ensure sufficient booking history. Similarly, listings that appeared inactive—those with a last review date within the past year—were also filtered out.

2. Data formatting

Raw data for host response rate and host acceptance rate were recorded as percentages (with % sign) and price with a dollar sign. All 3 variables are stripped to retain only the numeric value.

The price variable in the raw dataset is recorded as a string containing the “$” symbol. It was converted to numeric format for analysis.

3. Missing Data

To ensure that the study takes recency into consideration, only active listings (estimated occupancy of more than 0) are taken into consideration. This filtering allows the study to focus on listings that are currently engaged on Airbnb to better represent factors influencing attractiveness and demand. 5213 rows dropped.

Listings with valid license number or marked as “Exempted” were coded to T (True) while listings with missing license information were coded as F (False). Host is Superhost is coded as “T” or “F”.Missing values are recoded as “F”

4. Feature Engineering for Amenities variable

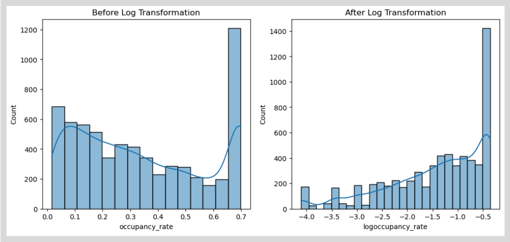
Original amenities were stored as long text strings (e.g., “Sulkin body soap”, “Treseme or Dove conditioner”). The data was cleaned and transformed into a count of amenities per listing.

5. Estimated occupancy rate

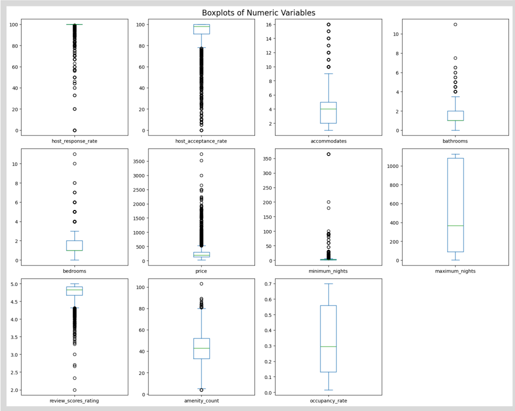
Derived by dividing estimated occupied days in the past year by 365 to obtain a continuous occupancy rate for multiple linear regression.

6. Variables transformation

Histogram of dependent variable and Boxplot of numerical values

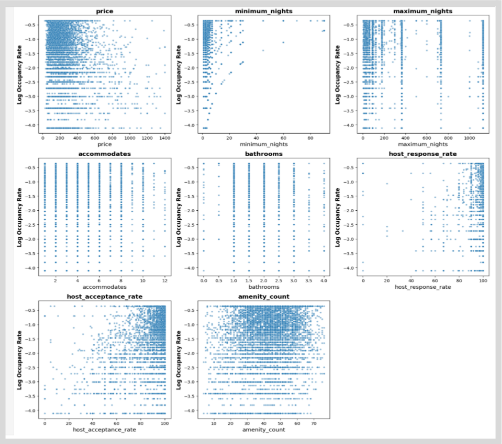


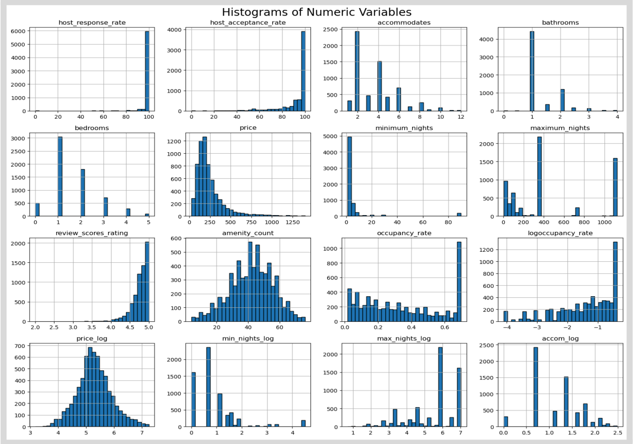
Our histogram for the dependent variable (occupancy rate) shows a right-skewed distribution with high concentration towards maximum at close to 250 nights. A logarithmic transformation is applied to reduce skewness and produce a distribution that is closer to normal.



Boxplot for numerical variables revealed the presence of extreme outliers. Such extreme values can distort model fit and bias coefficient estimates. A 99th percentile cut-off has been applied, retaining the vast majority of the data while reducing the impact of extreme outliers. This study focuses on active listings and extreme outliers are unlikely to represent typical market conditions. 10,065 rows remained after dropping outliers.

Pairplot (against log occupancy rate) and histograms of numerical values

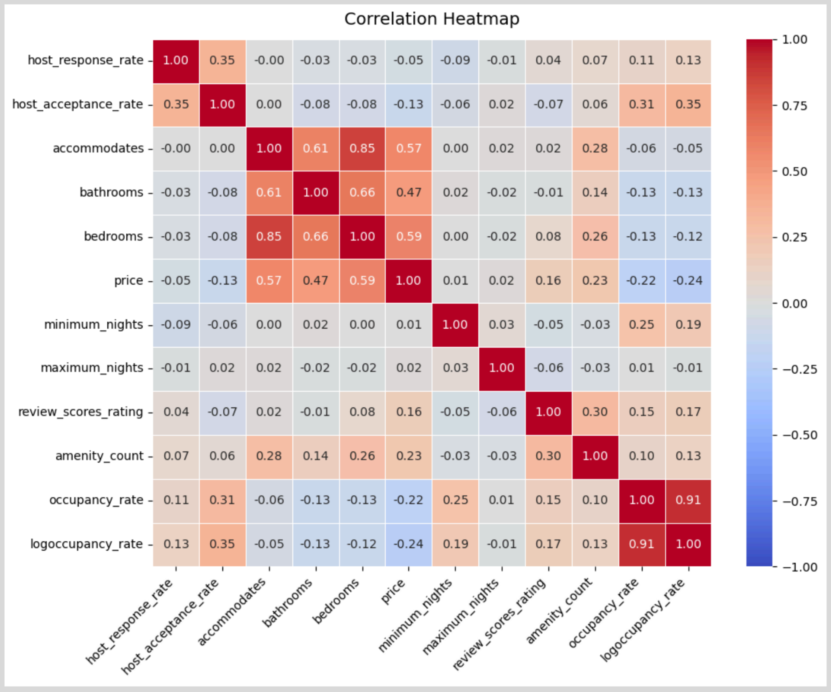




Pairplots and histograms for all numerical variables were plotted against the log-transformed estimated occupancy rate.

Price displayed a downward curve that flattened out suggesting that log occupancy rate tends to drop steeply when price increases and then levels off. Similarly, accommodates and minimum nights showed similar skewness suggesting that a few listings have disproportionately high values. Applying log-transformation can help to mitigate these effects.

7. Bedroom variable dropped



The correlation heatmap indicates a strong positive correlation (r=0.85) between bedrooms and accommodates. To reduce potential multicollinearity in the multiple linear regression model, bedrooms have been excluded. On the other hand the accommodates variable was retained as it is more representative of listing size.

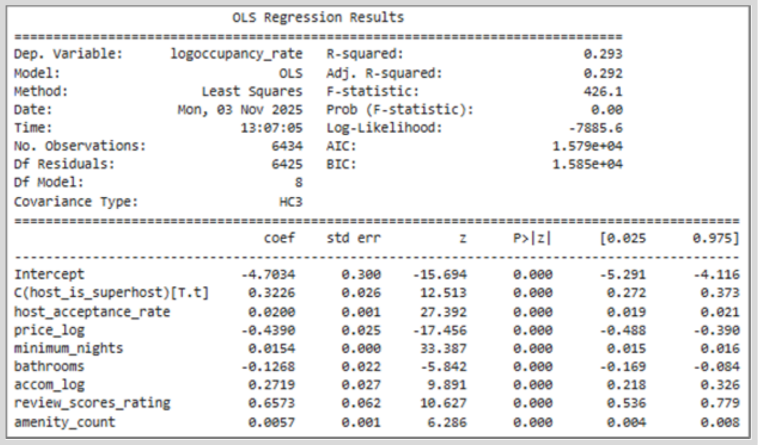
4. Methodology and Results

We used Multiple Linear Regression (MLR) as the primary analytical technique to understand which variable is most associated with *occupancy rate* from more than ten independent variables. Multiple linear regression is suitable because the dependent variable is continuous and influenced by a combination of quantitative and categorical explanatory variables. It is also important to note that after the data cleaning and filtering process, only 6434 listings remained.

4.1 Multiple linear regression

Our finalised multiple linear regression model is statistically significant with a F-statistic p-value below 0.00 and a modest adjusted R² score of 0.292 across 6434 listings. The explanatory variables used in the model were all statistically significant as well, with a p-value below 0.00.

Due to the limitations in available data, many factors affecting occupancy in the hospitality industry were not captured in our model. Our model focuses on listing characteristics that hosts are able to control. This is explained more in the Data Limitations section below.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:      logoccupancy_rate   R-squared:                       0.293
Model:                            OLS   Adj. R-squared:                  0.292
Method:                 Least Squares   F-statistic:                     426.1
Date:                Mon, 03 Nov 2025   Prob (F-statistic):               0.00
Time:                        13:07:05   Log-Likelihood:                -7885.6
No. Observations:                6434   AIC:                         1.579e+04
Df Residuals:                    6425   BIC:                         1.585e+04
Df Model:                           8
Covariance Type:                  HC3
=============================================================================================
                               coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------
Intercept                   -4.7034      0.300    -15.694      0.000      -5.291      -4.116
C(host_is_superhost)[T.t]    0.3226      0.026     12.513      0.000       0.272       0.373
host_acceptance_rate         0.0200      0.001     27.392      0.000       0.019       0.021
price_log                   -0.4390      0.025    -17.456      0.000      -0.488      -0.390
minimum_nights               0.0154      0.000     33.387      0.000       0.015       0.016
bathrooms                   -0.1268      0.022     -5.842      0.000      -0.169      -0.084
accom_log                    0.2719      0.027      9.891      0.000       0.218       0.326
review_scores_rating         0.6573      0.062     10.627      0.000       0.536       0.779
amenity_count                0.0057      0.001      6.286      0.000       0.004       0.008
```

Each coefficient represents the average change in occupancy rate associated with a one-unit change in that explanatory variable, holding **all other variables constant**.

Positive Associations

- **Review scores rating** (+0.657): Every 1 point increase in rating corresponds to 93% increase in occupancy on average
- **Superhost status** (+0.323): On average, Superhosts have approx 38% higher occupancy rate than non Superhost
- **Accommodates** (log) (+0.272): Every 1% increase in capacity (maximum number of guests) leads to 0.27% increase in occupancy on average
- **Host acceptance rate** (+0.02): Every 1% increase in host acceptance rate is associated with 2% increase in occupancy on average
- **Amenities count** (+0.0057): Increase in one amenity is associated with 0.57% increase in occupancy rate on average
- **Minimum nights** (+0.0154): Each additional minimum night is associated to 1.5% increase in occupancy on average

Negative Associations

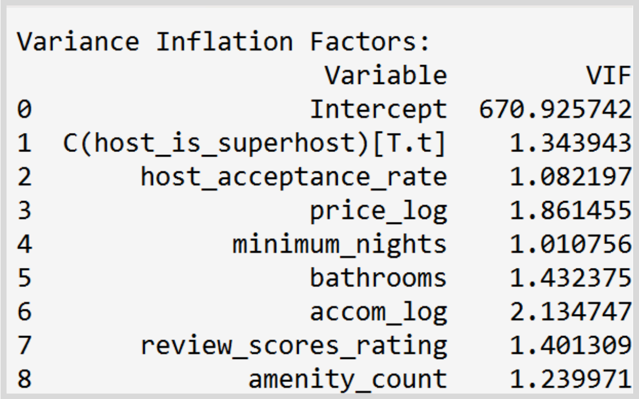
- **Price (log) (-0.439):** For every 1% increase in price decreases occupancy by 0.44% on average
- **Bathrooms (-0.127):** Every additional bathroom decreases occupancy by 13% on average

4.2 Technical Assumptions

To check the validity of our findings, we tested if our model met all the assumptions of linear regression. While our model met some assumptions, our regression had some issues with non-linearity and non-constant variance of residuals. As such, we applied the necessary treatments, which are explained in more details below.

1. Low/no multicollinearity between explanatory variables

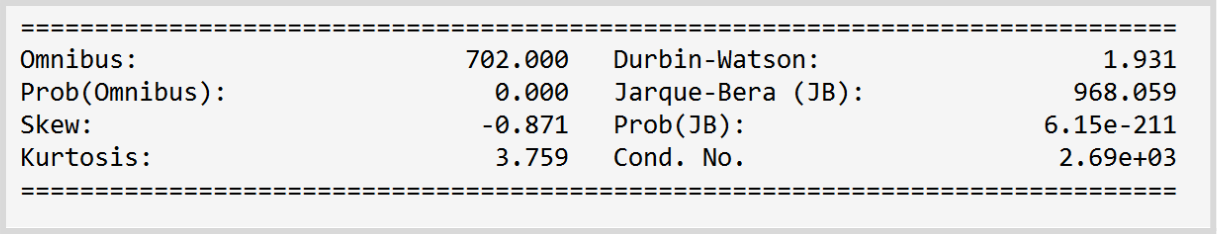
None of our explanatory variables had a Variance Inflation Factor (VIF) score above 10, indicating that there is no multi-collinearity.



```
Variance Inflation Factors:
                    Variable         VIF
0                  Intercept  670.925742
1  C(host_is_superhost)[T.t]    1.343943
2       host_acceptance_rate    1.082197
3                  price_log    1.861455
4             minimum_nights    1.010756
5                  bathrooms    1.432375
6                  accom_log    2.134747
7       review_scores_rating    1.401309
8              amenity_count    1.239971
```

2. Independence of residuals

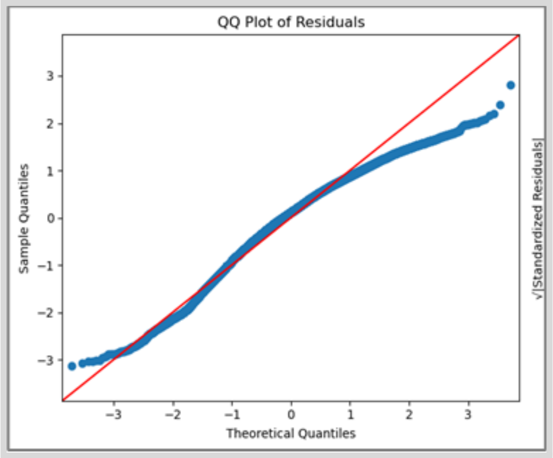
Our Durbin-Watson statistic of 1.931 falls between 1.5 and 2.5, which is the range that indicates normality of residuals.



```
==============================================================================
Omnibus:                      702.000   Durbin-Watson:                   1.931
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              968.059
Skew:                          -0.871   Prob(JB):                    6.15e-211
Kurtosis:                       3.759   Cond. No.                     2.69e+03
==============================================================================
```

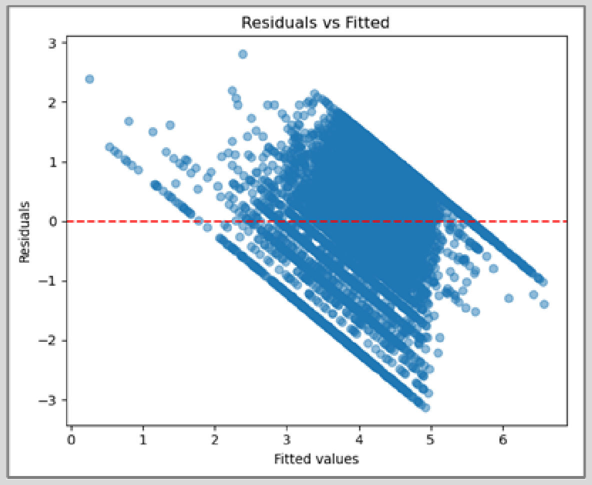
3. Normality of residuals assumption

As seen in our Q-Q plot, our residuals are normally distributed since most of the residuals fall along the line.



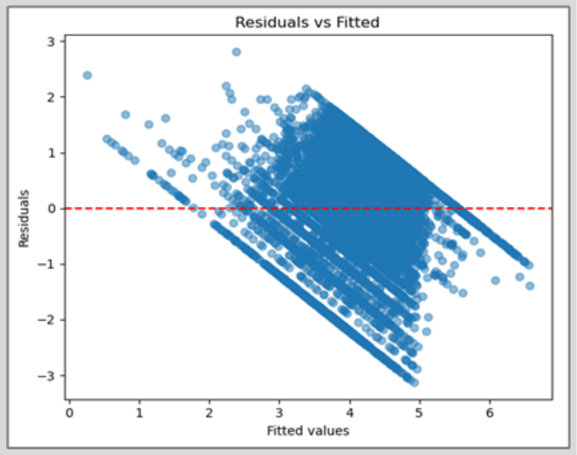
4. Homoscedasticity assumption

Our residuals versus fitted plot shows an issue of non-constant residual variance, since most of our data points are clustered into a band. To mitigate this issue, we applied robust standard errors to correct the standard-error estimates and obtain valid statistical inference despite the presence of heteroscedasticity.



5. Linearity

Our residuals versus fitted plot shows that the relationship between our explanatory variables and dependent variable (occupancy) was not linear, since most of our data points are clustered into a band. To mitigate this issue, we applied a logarithmic transformation to the dependent variable.



4.3 Data Limitations

While our adjusted R² score is modest at 0.292, we believe that this is largely due to the limitations in the data set. Specifically:

1. The dependent variable, occupancy rate, was derived from the number of reviews since occupancy data was unavailable in the original dataset.

This approach is aligned with other studies on the impact of Airbnb on the property market, as those studies used review-to-stay conversion rates as occupancy proxies. This indicates that the review-to-stay conversion is widely accepted, perhaps because Airbnb does not publish this data.

Furthermore, the reported conversion rates vary widely (31 % – 72 %) across studies, which is a large range and would affect the statistical models depending on the value used. For our project, we adopted a midpoint of 50%.

2. Due to our available statistical tests and data, we could not account for all variables relevant to occupancy

Airbnb has claimed that these variables could increase a listing's occupancy**:**

- High-quality photos
- Fact-based, precise listing description
- Desirable amenities, such as a pool, wifi, hot tub and air-conditioning
- Dynamic pricing strategy (e.g. weekly or monthly discount)
- Professional response to guest reviews, especially negative reviews

However, data for these variables are not included in most public Airbnb datasets, likely due to the difficulty of including said variables into the dataset. For example, most data sets would not include the analysis of photos of a listing, and whether they are classified as "high quality".

5. Management Recommendations

The model shows that a one-point increase in review score rating corresponds to a 93% rise in occupancy. These findings are supported by research by Xie et al. (2014), which found that better reviews for accommodations are associated with higher occupancy rates. Hence, based on the results, there is sufficient evidence to support Airbnb's investment in host support systems. Airbnb can introduce a host effectiveness dashboard that lets hosts track their reviews. The dashboard will provide personalised recommendations to ensure continuous improvement in service standards.

Another important insight is that Superhost status was one of the strongest positive drivers of occupancy, with Superhosts achieving 38% higher booking rates than non-Superhosts. This showcases that guests value trust and credibility when selecting a listing. Airbnb can provide awareness of the benefits of achieving Superhost status and the support required to do so. For example, Airbnb can advertise the benefits of Superhosts through push and in-app notifications to current hosts, as well as their progress to achieving Superhost status.

Our model also showed that a 1% price increase corresponds to a 0.44% drop in occupancy on average. Airbnb can develop a pricing machine learning model based on past data and provide hosts with suggested prices based on predictive modelling. This provides hosts with the flexibility to understand the most suitable pricing for their listing. Additionally, to overcome seasonal price fluctuations, Airbnb can also provide alerts to notify hosts of upcoming events that may lead to travel decline or spikes. Consequently, hosts would be able to adjust rates in advance and with confidence.

6. Conclusion

This project's objective was to identify the listing attributes most strongly associated with Airbnb occupancy in Sydney. The Multiple Linear Regression (MLR) provided Airbnb with clear, data-driven insights to optimise host listings. By analysing 6434 active listings, the top three variables that were highly influential are superhost status, price, and review score. Based on the findings, there is sufficient evidence to suggest that trust and service quality are significant indicators of high occupancy.

Due to data limitations, our MLR can only explain a moderate proportion of the occupancy variance. However, it is critical to note that occupancy is highly volatile and can fluctuate for numerous reasons. Future research can include photo quality or quality of host responses to strengthen findings.

In conclusion, this report provides an explanatory and analytical perspective on the variables most strongly associated with Airbnb occupancy. To ensure a more accurate interpretation, this report also considered the data limitations to provide a more realistic understanding of our objective. Based on our model's results, we proposed some recommendations that Airbnb can utilise to enhance listing performance. In a dynamic and digital environment, data-driven improvements are the core to sustaining a competitive advantage. By implementing these recommendations Airbnb can strengthen revenue strategies and equip hosts with actionable insights to optimise their listings.

References

Airbnb *Resource Centre*. (2024). *Help a host's listing stand out.* https://www.airbnb.com.sg/resources/hosting-homes/a/help-a-hosts-listing-stand-out-633?_set_bev_on_new_domain=1762137416_EAMDMxN2JkNDI1ZG

Chen, Y., & Schuckert, M. (2016, July). Why people choose Airbnb over hotel? Paper presented at the 80th TOSOK Gangwon Pyeongchang International Tourism Conference, Pyeongchang, Gangwon Province, South Korea. https://www.researchgate.net/publication/305225839_Why_people_choose_Airbnb_over_Hotel

Kumar, Naveen. "Airbnb Statistics 2025: Users & Revenue Data." *Demand Sage*, 27 Dec. 2024, www.demandsage.com/airbnb-statistics/. Accessed 1 Oct. 2025.

Inside Airbnb (n.d.). – Data assumptions.
https://insideairbnb.com/data-assumptions/

United Nations Tourism. (2025). *UN Tourism World Tourism Barometer — International tourism up 5% in first half of 2025 despite global challenges.* UN Tourism. https://www.untourism.int/un-tourism-world-tourism-barometer-data

Xie, K. L., Zhang, Z., & Zhang, Z. (2014). *The business value of online consumer reviews and management response to hotel performance.* International Journal of Hospitality Management, 43, 1-12.