Inside the Stay: Understanding What Drives Airbnb Host Average Daily Bookings

Problem, Motivation and Objectives

The Problem

- The growth of short-term rental platforms like Airbnb have transformed the hospitality landscape.
- Identifying factors affecting revenue is essential for making decisions on pricing, property features, and listing strategies.

Motivation and Objectives

- Airbnb hosts in Amsterdam encountered challenges in maintaining steady profit growth due to fluctuating travelling patterns and growing competition.
- Revenue on Airbnb is shaped primarily by two key variables, which are occupancy rate and listing price.
- Our study aims to investigate the key driving factor occupancy rate using the variable **average daily bookings** for Amsterdam Airbnb hosts to achieve and maintain revenue in a tightening market.

Analytical Questions: What are the main factors that are associated with the average daily bookings of accommodations?

- How do property attributes (e.g., room type) associate with average daily bookings?
- To what extent do host characteristics (e.g., Superhost status, acceptance rate) correlate with booking performance?

The Data

This project will be utilizing data on Airbnb listings in Amsterdam, The Netherlands as of 17 June 2025 downloaded from insideairbnb.com.

The data contains 79 columns and 10,168 rows with each record being a separate Airbnb listing in Amsterdam.

Key variables used for analysis during the study

Variable	Description	Data Type	Use
average_daily_bookings	Average number of bookings the listing receives per day	Numerical	Represents booking performance and is used to assess how different factors affect average daily bookings.
estimated_occupancy_l365d	Estimated occupancy over the past 365 days	Numerical	To calculate variable *"average_daily_bookings"*.
host_since	Date the host joined Airbnb	Numerical	
host_is_superhost	Whether the host is a Superhost	Boolean	To investigate how these listing features and host attributes associate with average number of daily bookings.
host_acceptance_rate	Percentage of booking requests the host accepts	Numerical	
room_type	Type of room offered	Categorical (Nominal)	
host_response_time	Description of typical time taken by host to respond to guest's messages	Categorical (Ordinal)	
Instant_bookable	Whether the accommodation can be instantly booked	Boolean	

Data Dictionary

Assumptions & Limitations

- **Data Source:** The data for this analysis was collected by crawling the Airbnb front-end webpage, not from direct individual host backend access.
- **Limitation:** This method is expected to produce some variance when compared to the true transaction data. However, the data will still serve as a reliable approximation of the overall Airbnb performance in Amsterdam.
- **Scope:** Listings that were **delisted** before the data was scraped are not reflected in the dataset. This should not be a significant issue, as the data is considered representative of the population.
- **Dependent Variable (Y):** The '**average_daily_bookings**' variable is dependent on the underlying data based on the assumptions outlined in the table. As a result of this, we are unable to use number of reviews as part of the predictor to avoid multicollinearity, and this is believed to be a huge explanatory factors.

Variable	Derivation
estimated bookings	• 50% of the received reviews were utilized
Average length of stay	• The primary value used is the mean average length of stay for that specific city. • If the city average is not available, a default value of 3 nights per booking is used. • After this value is selected (either the city average or the 3-night default), it is compared to the listing's 'minimum nights' requirement. If the listing's minimum nights requirement is higher, that value is used instead.
estimated_occupancy_l365d	• **estimated bookings * Average length of stay**

https://insideairbnb.com/data-assumptions/

Overall Analytical Approach

- Initial EDA and data sensing performed to understand the data relationship
- Understand data variables
- Identify data limitations and constraint
- Identify outlier and missing Value

Correlation Matrix
- To identify high correlation
- To identify variables that will be useful in explaining average daily booking

Descriptive Statistics
- Visualisation to understand data nature and relations.

Log Transformation
- Log transformation are performed on high skew data.

Model Fitting & Variable Evaluation
- Evaluation of model statistics significance, variable t-test p value.

Assumption Check
- Validation of final mode and residual to ensure that model assumptions are met.

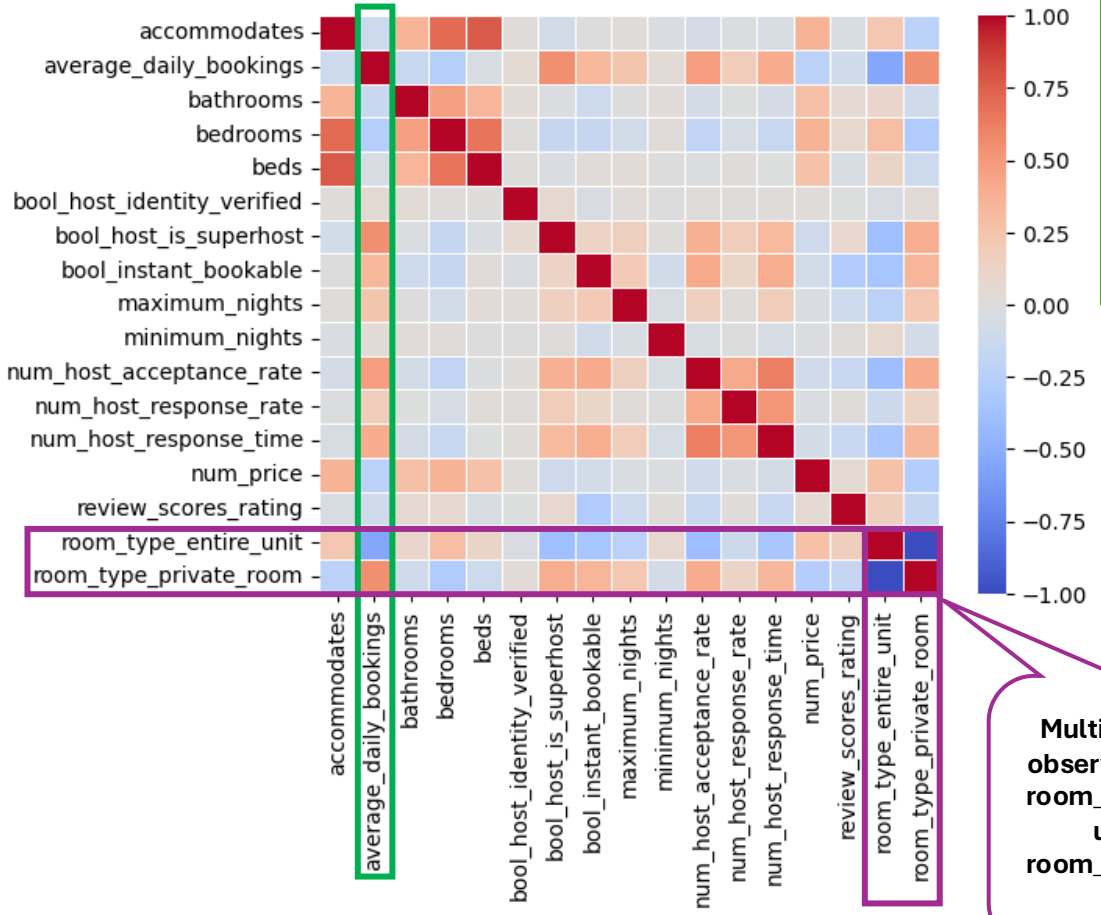
Data Preparation

1. **Handle Missing Values (Drop NA)**
   - Removed records containing missing values in essential variables to ensure model reliability.
   - This reduced noise and prevented bias during model fitting.
2. **Type Conversion → Numeric**
   - Converted percentage and currency strings (e.g., host_acceptance_rate, host_response_rate, price) to numeric types
   - Ensured data consistency and compatibility with regression modelling.
3. **Dummy Encoding**
   - Transformed categorical variables into dummy variables for model input.
   - Example: room_type_Entire home/apt, room_type_Private room, etc.
4. **Target Variable Creation**
   - Created a new variable representing the **listing's estimated occupancy** (based on assumptions from Inside Airbnb data documentation).
   - Used this as the dependent variable for explanatory modelling.
5. **Category Filtering (< 30 Samples)**
   - Excluded underrepresented categories (e.g., rare room types) with fewer than 30 observations.
   - Improved model reliability and reduced the impact of noise from sparse data.
6. **Removed Extreme Outliers**
   - Filtered listings with unrealistic revenue values (above $110,000 annually).
   - Maintained data quality by focusing on typical host behaviors and preventing skewed regression coefficients.

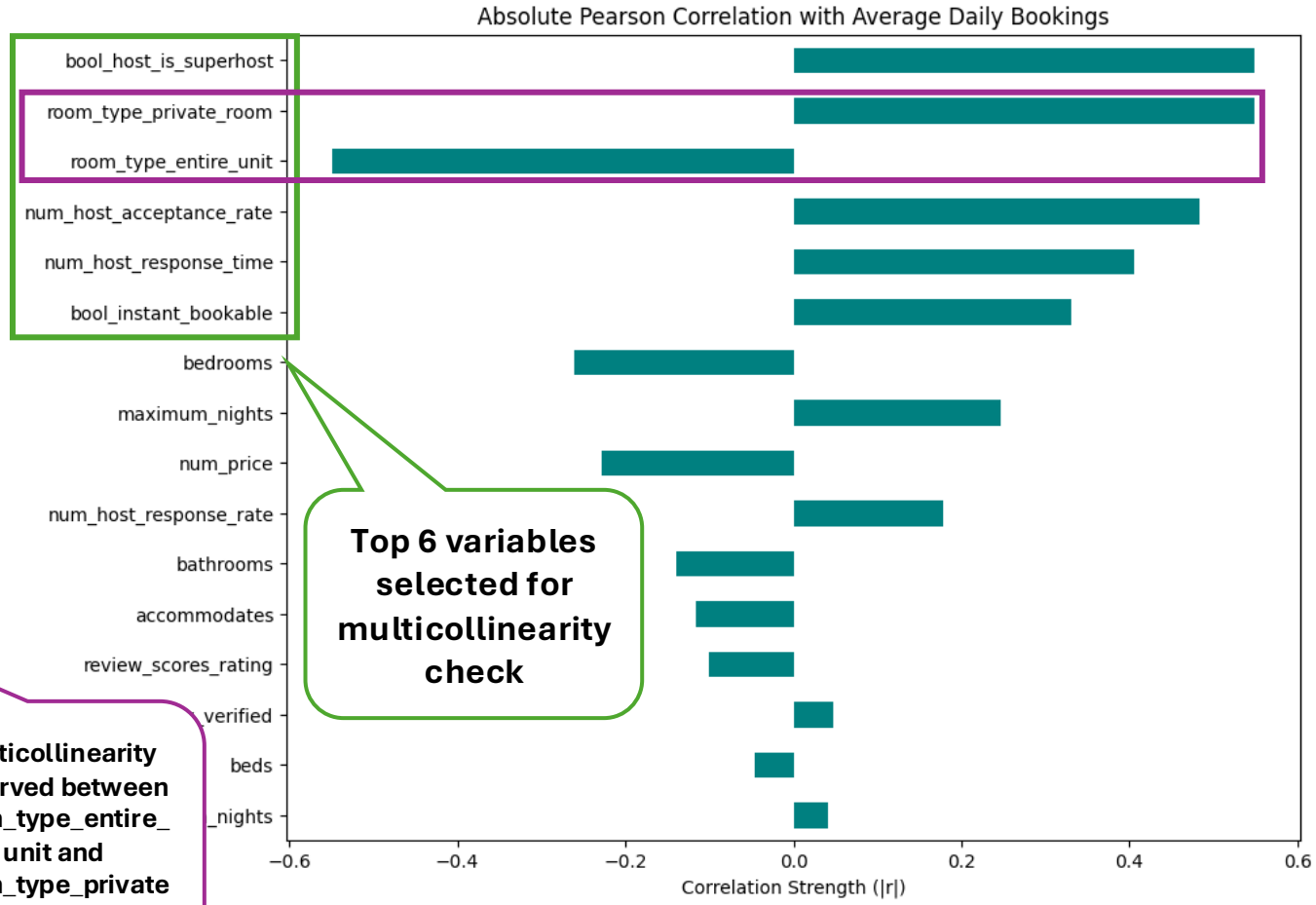
Correlation Matrix & Regression Model Setup

Objective: 1) To shortlist potential input / independent variables to be fitted into the MLR model;
2) To observe multicollinearity in selected input / independent variables

Correlation Matrix on 'average_daily_bookings'



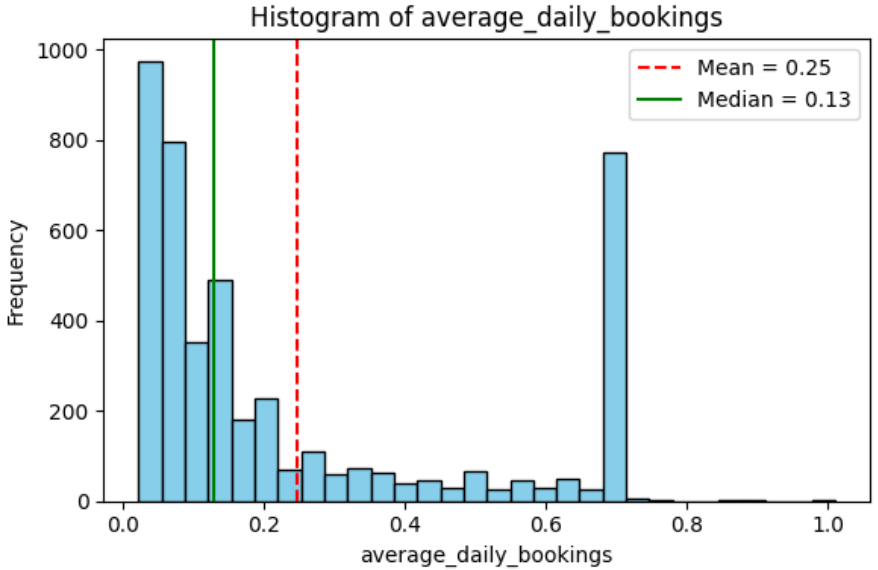
Multicollinearity observed between room_type_entire_unit and room_type_private_room



Top 6 variables selected for multicollinearity check

Descriptive Statistics

Target / Dependent Variable

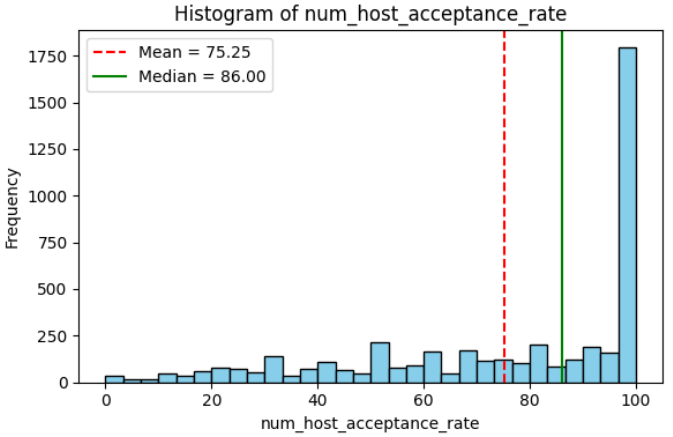


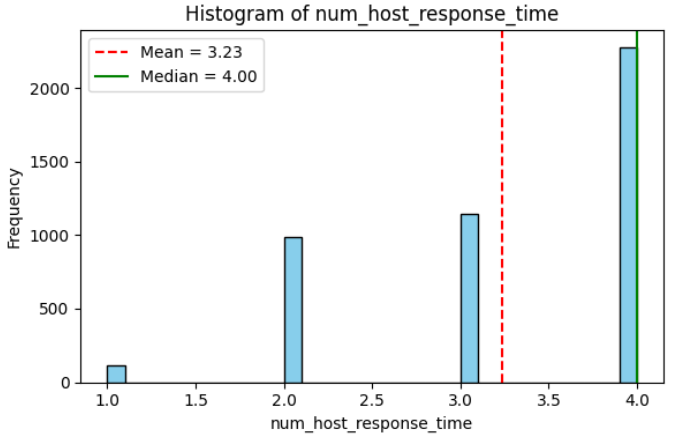
- **Mean (0.25) > Median (0.13)**, i.e. distribution is **right-skewed** (**positively skewed**).
- **Most listings** have **low daily bookings** (0.0 – 0.2).
- A **small group** of listings show **much higher booking rates** (around 0.7).

Interpretation:

- Most Airbnb listings are **under-booked**, suggesting **potential to raise occupancy**.
- A **distinct group of high-performing listings** exist, worth further analysis.

Numerical Input / Independent Variables

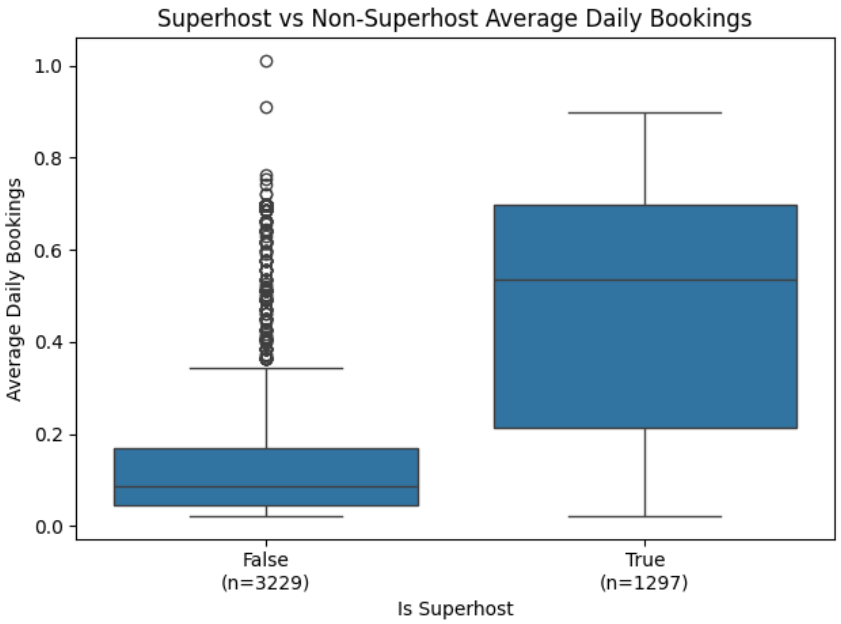


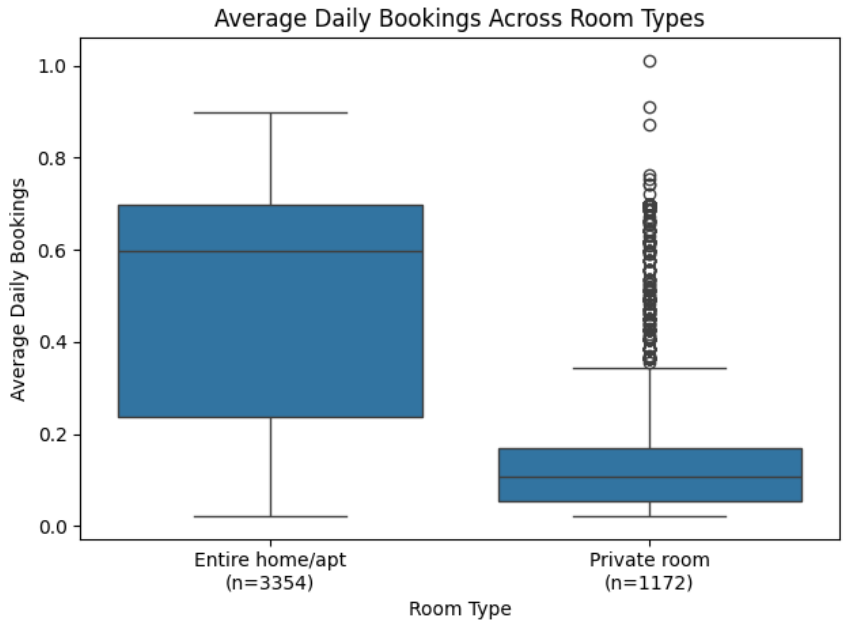


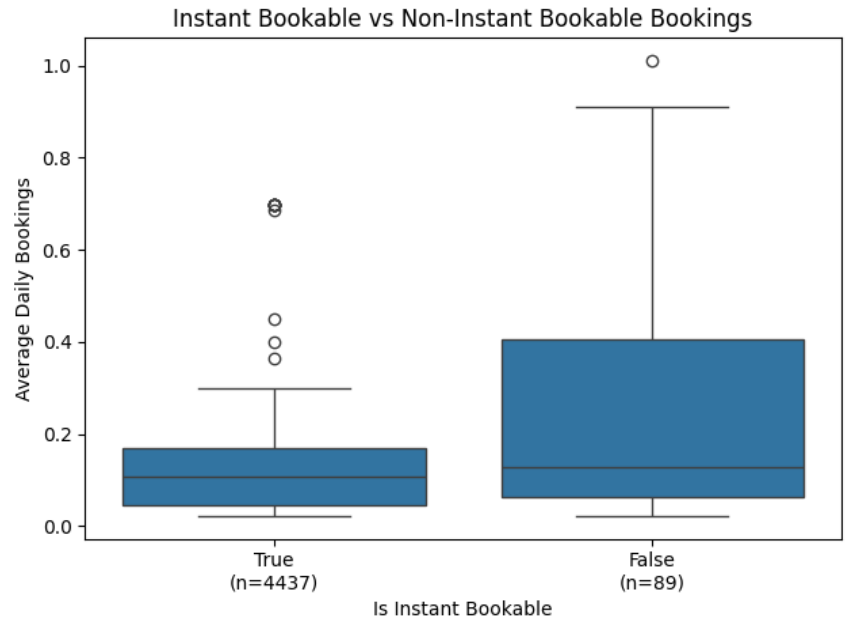
Variable	Acceptance Rate	Response Time
Skewness	Left (negatively skewed)	Left (negatively skewed)
Mean	75.25	3.23
Median	86.00	4.00 (within an hour)
Observations	Most hosts have very high acceptance rates (>80%)	Most hosts respond within an hour
Interpretation	It might be difficult for motivated hosts to improve further on this aspect	It might be difficult for motivated hosts to improve further on this aspect

Descriptive Statistics

Categorical Input / Independent Variables



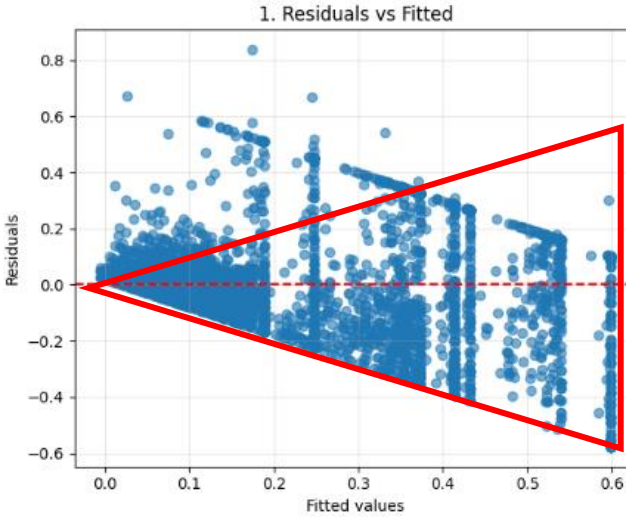




Variable	Median	Distribution	Observation	Interpretation
Superhost	**Median = 0.5 (Superhost)** vs **0.1 (Non-superhost)**	**Superhosts** show **wider IQR** while **non-Superhosts** have **more outliers**	Superhosts achieve higher but more varied booking performance	**Superhosts** experience **stronger demand and visibility** likely due to host perceived reliability
Room Type	**Median = 0.6 (Entire home/apt)** vs **0.1 (Private room)**	**Wider IQR** for **entire homes** but **more outliers** for **private rooms**	Entire units achieve higher but more varied booking performance	Listings **offered as an entire home / apartment tend to do better** in terms of bookings as compared to shared spaces
Instant Bookable	**Median = 0.1 (Non-instant and Instant)**	**More outliers** for **instant bookable** listings but **wider IQR** for **non-instant bookable**	Non-instant bookable has more varied booking performance	Despite having a larger variation, there are **some unique non-instant bookable listings** that **perform well**

Multiple Linear Regression (Baseline)

Assumption Check

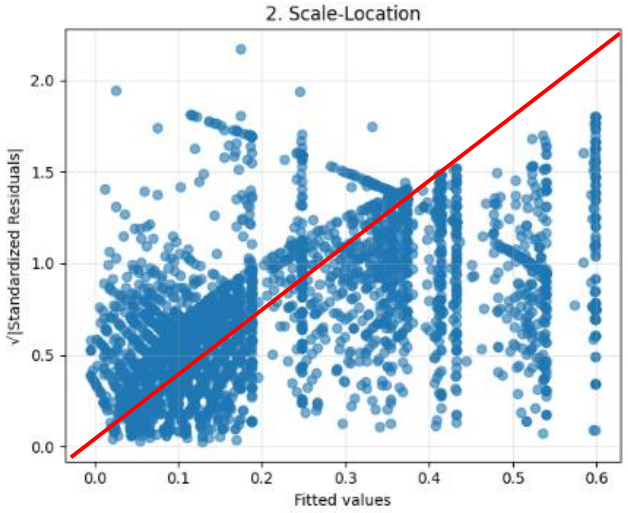


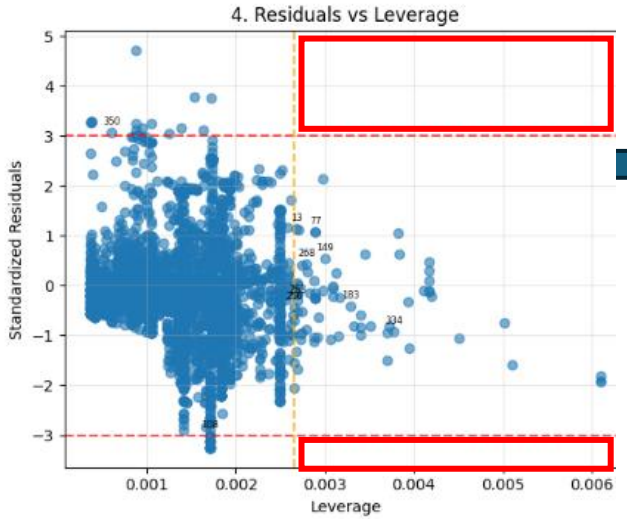
1. Residuals vs. Fitted

-Residuals show **random scatter**: **linear relationship** confirmed
-**'Funnel' shape** noted: **possible heteroscedasticity**

2. Scale-Location

-Standardized residuals show an **upward trend, i.e. positive trend**
-Scatter is **not randomly horizontal**
-**Possible heteroscedasticity**



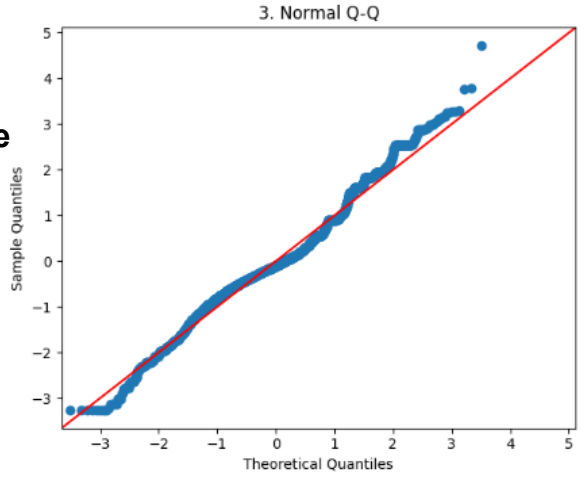


4. Residuals vs. Leverage

-Zero observations with both high leverage and high residuals.
-No **high-leverage outliers** detected.

3. Q-Q Plot

-Points **align closely along the diagonal line**.
-**Near normal residuals** with slight tail deviations.

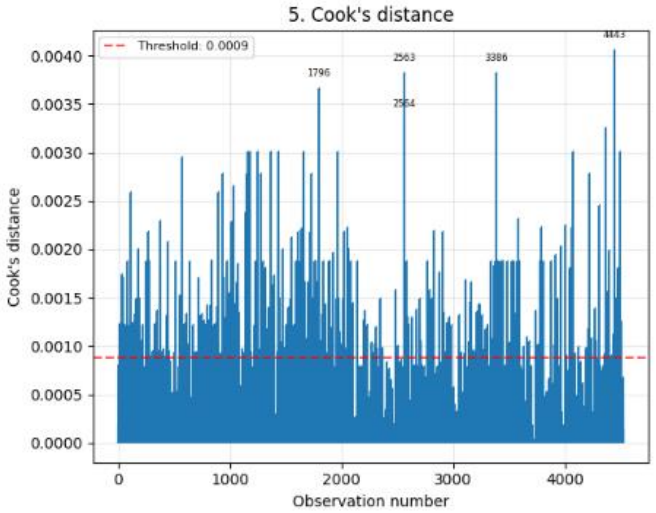


5. Cook's Distance

High influential point count - **450 strong influential points**

Max Cook's D = 4.6 × threshold

Potential model instability



<u>Solution</u>: To perform log transformation of input / dependent variable Y (average_daily_bookings)

Multiple Linear Regression

Model Transformation & Refitting

Before Transformation & Refitting

OLS Regression Results

Dep. Variable:	average_daily_bookings	R-squared:	0.482
Model:	OLS	Adj. R-squared:	0.481
Method:	Least Squares	F-statistic:	839.8
Date:	Mon, 03 Nov 2025	Prob (F-statistic):	0.00
Time:	04:49:53	Log-Likelihood:	1403.0
No. Observations:	4526	AIC:	-2794.
Df Residuals:	4520	BIC:	-2755.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0215	0.011	-1.969	0.049	-0.043	-9.21e-05
bool_host_is_superhost	0.1852	0.007	27.835	0.000	0.172	0.198
room_type_private_room	0.1660	0.007	23.384	0.000	0.152	0.180
num_host_acceptance_rate	0.0014	0.000	11.100	0.000	0.001	0.002
num_host_response_time	0.0165	0.004	4.197	0.000	0.009	0.024
bool_instant_bookable	0.0592	0.008	7.652	0.000	0.044	0.074

Omnibus:	152.490	Durbin-Watson:	1.806
Prob(Omnibus):	0.000	Jarque-Bera (JB):	244.197
Skew:	0.308	Prob(JB):	9.40e-54
Kurtosis:	3.956	Cond. No.	356.

Log transformed target variable

After Transformation & Refitting

OLS Regression Results

Dep. Variable:	log_average_daily_bookings	R-squared:	0.426
Model:	OLS	Adj. R-squared:	0.425
Method:	Least Squares	F-statistic:	1117.
Date:	Mon, 03 Nov 2025	Prob (F-statistic):	0.00
Time:	06:10:37	Log-Likelihood:	-5596.2
No. Observations:	4526	AIC:	1.120e+04
Df Residuals:	4522	BIC:	1.123e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.1040	0.037	-84.877	0.000	-3.176	-3.032
bool_host_is_superhost	0.8158	0.031	26.400	0.000	0.755	0.876
room_type_private_room	0.6374	0.032	19.742	0.000	0.574	0.701
num_host_acceptance_rate	0.0101	0.001	19.945	0.000	0.009	0.011

Omnibus:	164.195	Durbin-Watson:	1.790
Prob(Omnibus):	0.000	Jarque-Bera (JB):	189.267
Skew:	-0.443	Prob(JB):	7.96e-42
Kurtosis:	3.466	Cond. No.	244.

2 variables removed from final MLR model due to:

1. Relatively low t-value
2. Relatively low net coefficient value

3 variables retained in final MLR model

Multiple Linear Regression

Model Results

Model Evaluation

Variable	Value
R-squared	0.426
Adj. R-squared	0.425
Prob (F-statistic)	0.00
BIC	1.123e+04

Model captures meaningful patterns in booking behavior.

Highly statistically significant model. Model reliably explains relationships beyond random chance.

Model represents efficient trade-off between accuracy and simplicity.

Result Interpretation

Variable	coef	std err	T - P>|t|	[0.025	0.975]
const	-3.1040	0.037	-84.877 - 0	-3.176	-3.032
Super-host	0.8158	0.031	26.400 - 0	0.755	0.876
Acceptance rate	0.0101	0.001	19.945 - 0	0.009	0.011
Room-type private-room	0.6374	0.032	19.742-- 0	0.574	0.701

4. Confidence Interval Analysis
Narrow confidence intervals reflect high estimation precision. True population parameters are reliably captured within tight bounds.

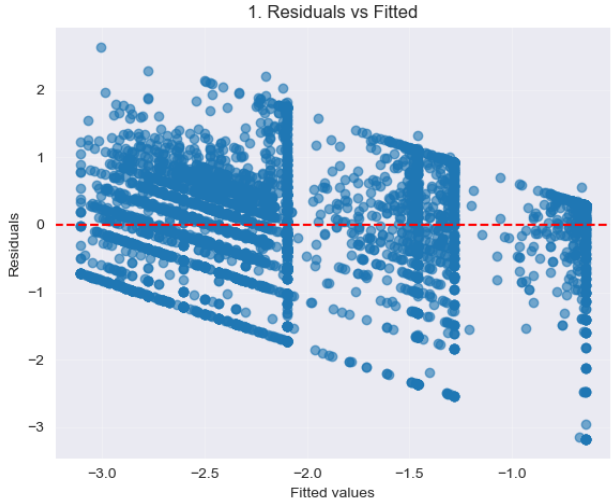
3. Statistical Significance & Confidence
All t-statistics > **19.7** and p-values = 0.000. We reject null hypotheses with high confidence for all variables.

2. Statistical Precision & Reliability
High precision across all estimates with small standard errors.

1. Coefficient Estimates & Economic Significance
All variables show strong positive effects on bookings. Superhost status has the largest effect on daily bookings.

Multiple Linear Regression (Final)

Assumption Check

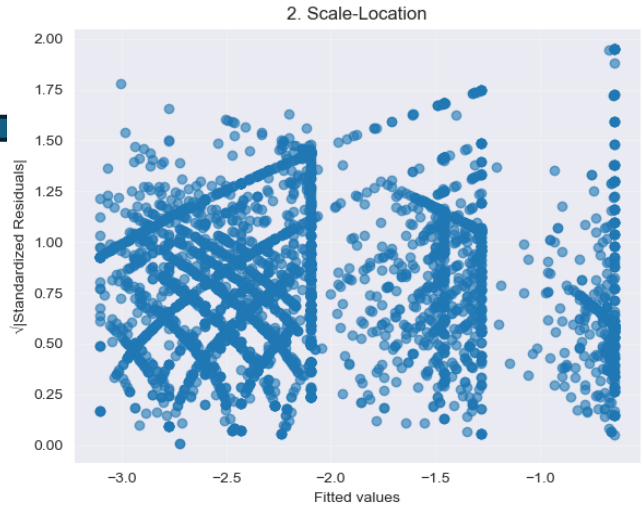


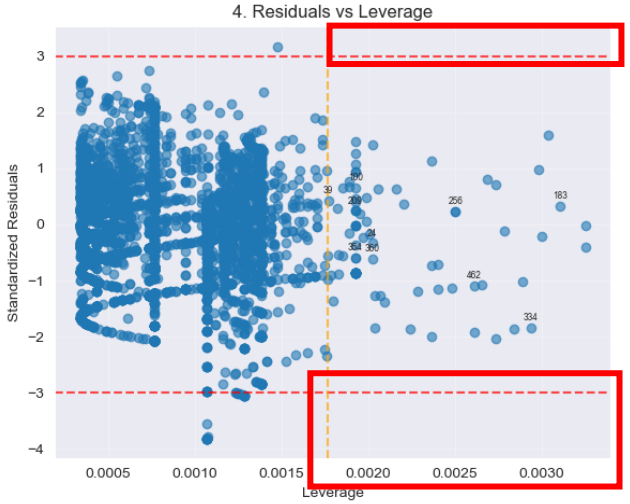
1. Residuals vs. Fitted

-Residuals show **random scatter**
-Linear relationship confirmed

2. Scale-Location

-Standardized residuals randomly distributed with **no trend**.
-No obvious heteroscedasticity after log transformation.



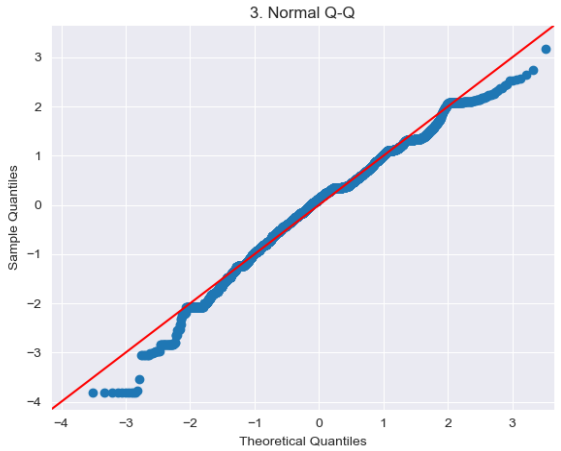


4. Residuals vs. Leverage

-Zero observations with both high leverage and high residuals.
-No **high-leverage outliers** detected.

3. Q-Q Plot

-Points align closely along the diagonal line.
-Near normal residuals with slight tail deviations.

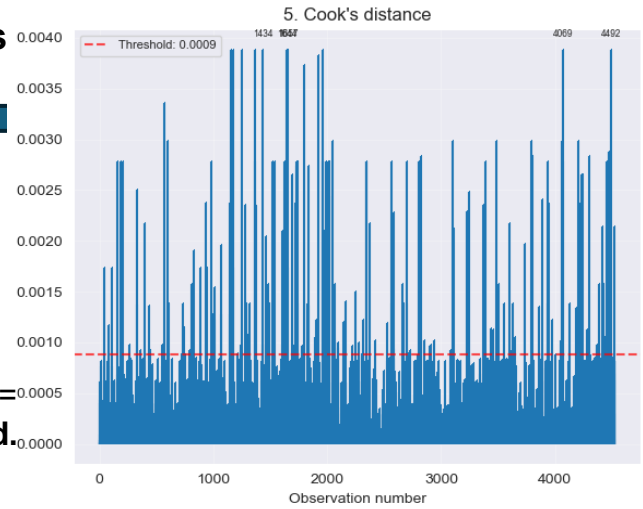


5. Cook's Distance

Low influential point count **(3.7%).**
166 strong influencers

max Cook's D = **4.4× threshold.**

Coefficient estimates stable; model is reliable.



Business Recommendations

1. Build Trust and Credibility as a Superhost

- Strive to achieve and maintain Superhost status as it's one of the clearest indicators of guest trust and high-quality hosting.
- Be proactive: respond quickly, ensure a spotless stay, and handle guest issues professionally.
- A strong reputation not only increases visibility in search results but also makes guests more confident which drives higher booking conversions.
- Consistency is key, every 5-star stay brings you closer to long-term success.

2. Improve Responsiveness and Booking Acceptance

- Keep your acceptance rate high by avoiding unnecessary declines or cancellations.
- Make booking easy for guests: keep your calendar accurate, offer flexible stay options, and consider enabling Instant Book.
- Guests prefer hosts who respond quickly as it creates a smoother experience and signals reliability.
- Every accepted booking increases your chance of positive reviews and future visibility.

3. Differentiate Your Space Through Experience

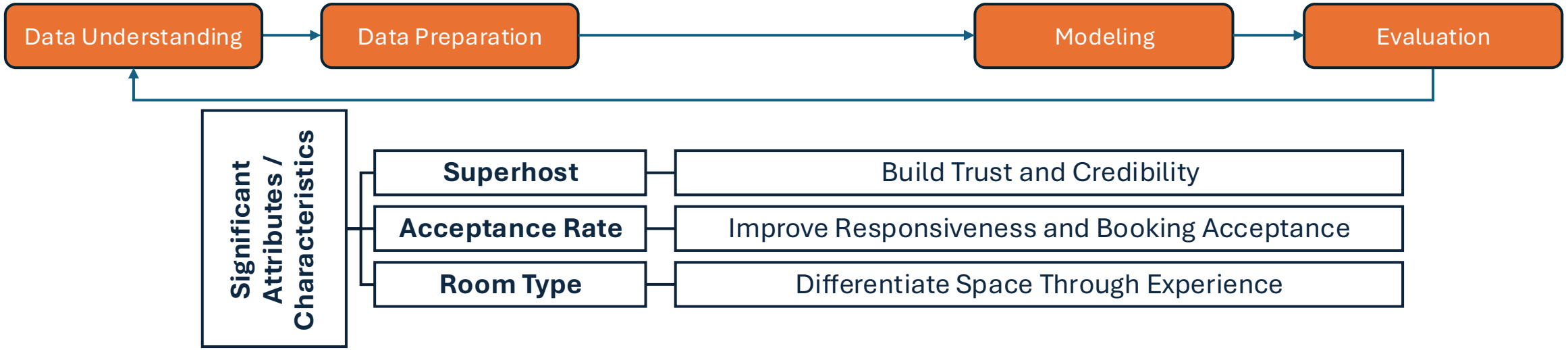
- Whether it's an entire home or a private room, focus on what sets your listing apart.
- For private-room hosts: emphasize authentic local experiences, comfort, and affordability to attract travelers seeking value and connection.
- For entire-home hosts: highlight privacy, space, and convenience while maintaining the warmth of a home environment.
- Guests remember experiences more than features because thoughtful touches and good communication can turn first-time visitors into repeat guests.

Conclusion & Future Work

Recap Analytical Questions:

How do property attributes (e.g., room type) associate with average daily bookings?

To what extent do host characteristics (e.g., Superhost status, acceptance rate) correlate with booking performance



Future Work

- **Model Enhancements:** Progress from explanatory to predictive modelling and recommended pricing / attributes for hosts
- **Data Improvements:** To obtain the true backend data to perform better analysis since dependent variable is currently estimated. To additionally obtain timing / seasonal data which enables further analysis based on tourism season / timing.
- **Sentiment Analysis:** To perform sentiments analysis on text fields, e.g. review content, neighbourhood description, etc.