EBA5002: Graduate Certificate in Business Analytics Practice

Data-Driven Insights to Enhance Airbnb Occupancy

Date of Report

27 Apr 2025

Table of Contents

Executive Summary

This report outlines Airbnb-lytics' approach to providing value to SingStay (a fictitious local short-term property rental company). With the growing demand in the short-term rental market in Singapore, SingStay aims to enhance its Airbnb platform listings to increase occupancy rates through technology-driven data analysis.

Airbnb-lytics adopted a hybrid implementation framework and analytics workflow based on CRISP-DM. This approach ensures robust project governance, increases solution scalability and adaptability, and improves customer satisfaction by incorporating feedback collected during short sprint cycles.

This project utilizes the publicly available "listings.csv" and "reviews.csv" from Inside Airbnb website. The project focuses on text and sentiment analysis to investigate trends in guest reviews and competitor descriptions. It builds predictive models to identify key attributes influencing occupancy and develop visualizations to monitor and continuously evaluate the listing performance of implemented strategies. To strategically improve listings, SingStay should incorporate the project findings and modeling solutions with their data for comparison, and develop a dashboard to continuously evaluate the results.

The report details the data governance, profiling, dimensional modeling, and exploratory data analysis steps undertaken. Finally, the report outlines the project's limitations and provides future recommendations to address them.

List of Figures

List of Tables

Introduction

Airbnb-lytics is providing business analytics solutions. The primary objective of this project is to address SingStay's low occupancy rates on Airbnb listings. To achieve this, the team will leverage technology-driven data analysis, utilizing the publicly available dataset from Inside Airbnb website ("listings.csv" and "reviews.csv") to identify the key success factors.

Background

Singapore's short-term rental market is a rapidly growing segment of the hospitality and tourism industry. This sector provides flexible lodging options for both leisure and business travelers seeking alternatives to traditional hotels. The global short-term vacation rental market was valued at USD 134.51 billion in 2024 and is projected to grow at a CAGR of 11.4% from 2025 to 2030 (Grand View Research, n.d.).

Despite this growth, businesses must remain competitive by optimizing their pricing strategies, amenities, and marketing efforts to attract guests. SingStay, a relatively new entrant in this space, has struggled to achieve sustainable occupancy rates.

Business Problem

Although SingStay has been operating in Singapore's expanding short-term rental market for a year, it continues to experience persistently low occupancy rates on Airbnb. This underperformance leads to revenue losses and higher operational costs, hindering the company's ability to compete effectively. With Singapore's tourism industry experiencing strong growth and increasing visitor arrivals (Renald Yeo, 2024), the demand for Airbnb accommodations is expected to rise. However, without strategic interventions, SingStay risks missing out on this opportunity.

To address this challenge, this project will apply data-driven analytical methods to study successful competitor listings, identify key attributes influencing high occupancy, and develop actionable recommendations. By leveraging these insights, SingStay can optimize its listing strategies, enhance booking performance, and strengthen its market position on Airbnb.

Key Objectives and Requirements

Business Objectives & Requirements

As a new entrant in short-term rental market experiencing persistently low occupancy rates on Airbnb, SingStay aims to leverage valuable insights from publicly available dataset from Inside Airbnb website to boost occupancy rates and optimise pricing strategies.

The three key business requirements are to:

- **Identify and prioritise key performance indicators (KPIs) to improve bookings.**
  Perform analysis of customer reviews to understand booking preferences and inform KPI determination.

- **Gain insights into competitor listings and strategies**
  Deep dive into listing data to uncover key features (e.g., location, Superhost status, pricing, etc.) that drive high occupancy rates.

- **Align business objectives by tracking performance of the adapt strategies**
  Develop visualisation of the identified KPIs that align with business goals enables SingStay to easily track performance, and adapt strategies.

Technical Objectives

The following technical objectives are transition from the above business requirements:

- **Text Analysis**
  To extract accurate and meaningful information from text data, this project employs a combination of text analysis techniques including topic modeling and sentiment analysis.

  Text analysis begins by performing word cloud of the tokens to provide a visual overview of the type of words that are prominent in text data, followed by Term Frequency-Inverse Document Frequency (TF-IDF) to filter out common words and bigram/trigram analysis to focus on word pairings that are more important of the text.

  Topic modeling identifies key themes in guest reviews and competitor listing descriptions. The process involves combining token columns, creating a dictionary and bag-of-words corpus (after filtering rare/common tokens), training an LDA model, saving the trained model, visualizing it with pyLDAvis, assigning topic distributions, and labelling the dominant topic per listing. The listings were labelled with the dominant topic to feed into the predictive modeling.

  Sentiment analysis is conducted using TextBlob dictionary to obtain the sentiment label to feed into a predictive model. In further iterations we will suggest SingStay to also evaluate the polarity and subjectivity of the text to understand the emotional tone of the guest from their reviews.

- **Predictive Models**
  Development of predictive model involved several key data preparation steps that includes merging topic modeling and sentiment analysis results, converting string data to numerical format (including one-hot encoding amenities and aggregating topic modeling counts), incorporating price ranges, and computing the estimated occupancy rate from feature/attributes 'availability_365' for a full-year perspective.

- **Visualisations**
  The predictive model identified top 15 features affecting occupancy, along with price consideration. Analysis insights are visually presented using bar charts, bubble chart, heatmap, and faceting, using the Matplotlib and Seaborn Python libraries.

Table 1.1: Business Objective, Technical Objective and Alignment to Business Objective

<table>
<tr><th>Business Objective</th><th>Technical Objective</th><th colspan="2">Alignment to Business Objective</th></tr>
<tr><td rowspan="2">Data Collection and Assessment: Gathering data on pricing trends, booking patterns, guest feedback, and property features.

Competitor Analysis: Identifying and analyzing successful Airbnb listings in the local market to determine the factors associated with high occupancy.</td><td>Build predictive models to identify key factors influencing occupancy.</td><td>Supports the business goal of improving occupancy rates by identifying and prioritizing factors (e.g., pricing, amenities) that drive bookings.</td><td rowspan="2">Helps maximize revenue by recommending dynamic pricing tailored to market conditions, ensuring a balance between occupancy and profitability.</td></tr>
<tr><td>Perform text analysis to investigate trends from guest reviews and competitor descriptions.</td><td>Enables a deeper understanding of competitors' key listing features and strategies, helping the company adapt to market demands effectively.</td></tr>
<tr><td>Optimization Recommendations: Developing tailored strategies to enhance the company's property listings, including pricing adjustments, improved marketing, and enhanced guest experience offerings.</td><td>Implement visualisation to monitor and evaluate changes in listing performance over time.</td><td colspan="2">Ensures consistent alignment with business objectives by tracking the effectiveness of improvements and adapting strategies as needed.</td></tr>
</table>

Project Plan

Project team has decided to adopt the Hybrid Implementation framework. It includes the Disciplined Agile Delivery during the Inception and Construction with analytics workflow based on CRISP-DM , and continuous improvement and deployment during the Transition & Rollout, and Maintenance Phase. (Tan, 2023)

Analytics Workflow

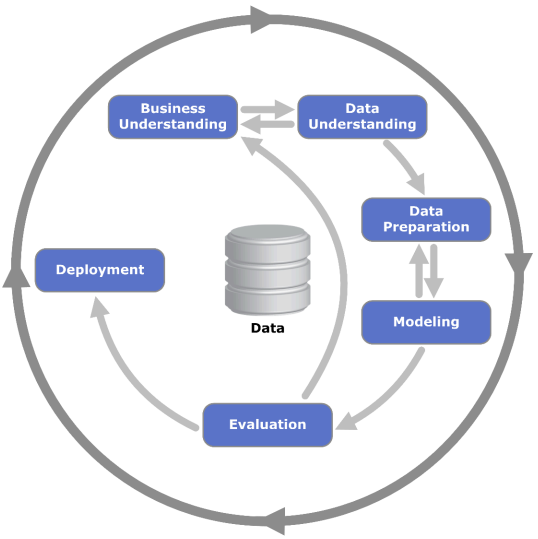


Figure 1.1: CRISP-DM (Almirgouvea, n.d.)

The analytics workflow follows the CRISP-DM methodology. The process begins with Business understanding, where objectives and project plans are defined and created. Data understanding involves data collection, exploration, and quality checks. Data Preparation focuses on data selection, cleaning, and transformation. The Modeling phase uses topic and predictive modeling techniques with model comparison and hyperparameter tuning. Evaluation assesses model performance against business goals, this includes the feedback compiled during the review and retrospective session with the stakeholder at the end of each sprint to enable iterative fine-tuning. The project concludes with Deployment, including testing and user training, and ongoing Monitoring and Maintenance that is outsourced to a 3rd party service provider.

Project Plan Optimization

As the project fell behind schedule during Sprint 1, the project team optimised the project plan. This involved analysing the project schedule to identify the critical path, and a column indicating the critical path was added to the Project Gantt Chart (link embedded). The burndown chart for this project can also be found in this link. Tasks were reviewed and reallocated to team members based on their skills and availability. Key stakeholders were informed of the deliverable reprioritisation, and consensus was reached to focus on delivering the most critical features. The scope and goals of subsequent sprints were adjusted to prioritise these critical features within the given timeline.

Stakeholders Management

SingStay and the project team have identified the key stakeholders and the available resources. The roles and responsibilities of these stakeholders are defined in the RACI chart below:

Table 1.2: RACI Chart

Project Deliverables		Executive & Project Sponsor	Business Users	Scrum Master	Product Owner	Analytics Team	IT Team
Scoping Phase	Identify business goals & requirements	A	R	R	R	I	C
	Stakeholders consensus	R	R	R	R/A	I	C
Planning & Design Phase	Define project schedule	I	C/I	R/A	C	C	C
	Exploratory Data Analysis	I	C	R	C	R/A	C
Developing Phase	Build Deliverables	I	C	R	R	R/A	C
	Change Management	I	C	R	A	R	C
	Testing	I	R	R	A	R	R/C
Deployment	Solution Adoption	A	R	R	R	C	R/C
Post Development	Support & maintenance handover document	I	I	R	A	R/C	R/C

Legend

R: Responsible **A**: Accountable **C**: Consulted **I**: Informed

IT Team: Data Engineers, System Engineers

Analytics Team: Data Analysts, Business Analysts, Developers

Risk Management

The project team has identified the potential risks, prioritised them using risk exposure methods, and outlined the risk mitigation plan in Table 1.3.

Table 1.3: Risk Management

Risk	Risk Description	Impact	Probability	Risk Exposure	Mitigation
Data Privacy & Security	Sensitive data that may be disclosed, and possible loss of data in transits	3	3	9 - High	Perform data masking, encryption, regular security audits, and implement access control strategy
Data Quality Issue	Inconsistent data that can lead to unreliable results	3	3	9 - High	Perform data validation, cleansing, automated quality checks, and establish data governance framework
Scope Creep	Informally including additional scope of work may result in an increase in project costs	3	3	9 - High	Define scope clearly, enforce change control processes, and maintain stakeholder alignment
Resource Availability	Assigned resources are not being able to perform the assigned tasks may result in project delays	3	3	9 - High	Develop resource plan, build buffer to allocate resource for key milestones of the project, get support and commitment from management and project team
User Adoption Challenges	Some users may resist using a new system or process.	2	2	4 - Medium	Conduct user training sessions, early-stage user engagement, and collect feedback during the Review & retrospective sessions

Definition of scale of Impact and Probability:

- Impact scale : 1 - Low; 2 - Medium; 3 - High
- Probability scale: 1 - Low; 2 - Medium; 3 - High
- Risk of Exposure: Impact x Probability

Project Scope

For this project, we will be using the “listings.csv” and “reviews.csv” datasets for Singapore obtained from the Inside AirBnB website.

Data Quality

The “reviews.csv” file contains a total of 39,282 rows and 6 columns. There are no duplicate rows in the dataset. The summary of the data can be found in the table as follows:

Table 2.1: Summary of “reviews.csv”

Data Feature	Data Type	Missing Values	Unique Values
listing_id	Integer	0	1,788
id	Integer	0	39,282
date	Date	0	3,902
reviewer_id	String	0	36,400
reviewer_name	String	0	16,835
comments	String	8	37,410

The “listings.csv file” contains a total of 3,381 rows and 75 columns. There are no duplicate rows in the dataset. The summary of the data for the key data features can be found in the table as follows:

Table 2.2: Summary of “listings.csv”

Data Feature	Data Type	Missing Values	Unique Values
id	Integer	0	3,381
host_response_rate	String	858	32
host_listings_count	Integer	0	43
neighbourhood	String	1,377	31
bathrooms	Float	807	19
has_availability	String	285	2
availability_30	Integer	0	31
availability_60	Integer	0	61
availability_90	Integer	0	90
availability_365	Integer	0	291
review_scores_rating	Float	1,593	125

For the data feature ‘availability_365’, it is also noted that there are 455 rows that are indicated as 365 available days. This may potentially suggest that the listings may no longer be active.

Dimensional Model

The schema is built around the star schema architecture, where a central fact table records measurable events (availability and occupancy), and multiple dimension tables provide context and descriptive attributes.

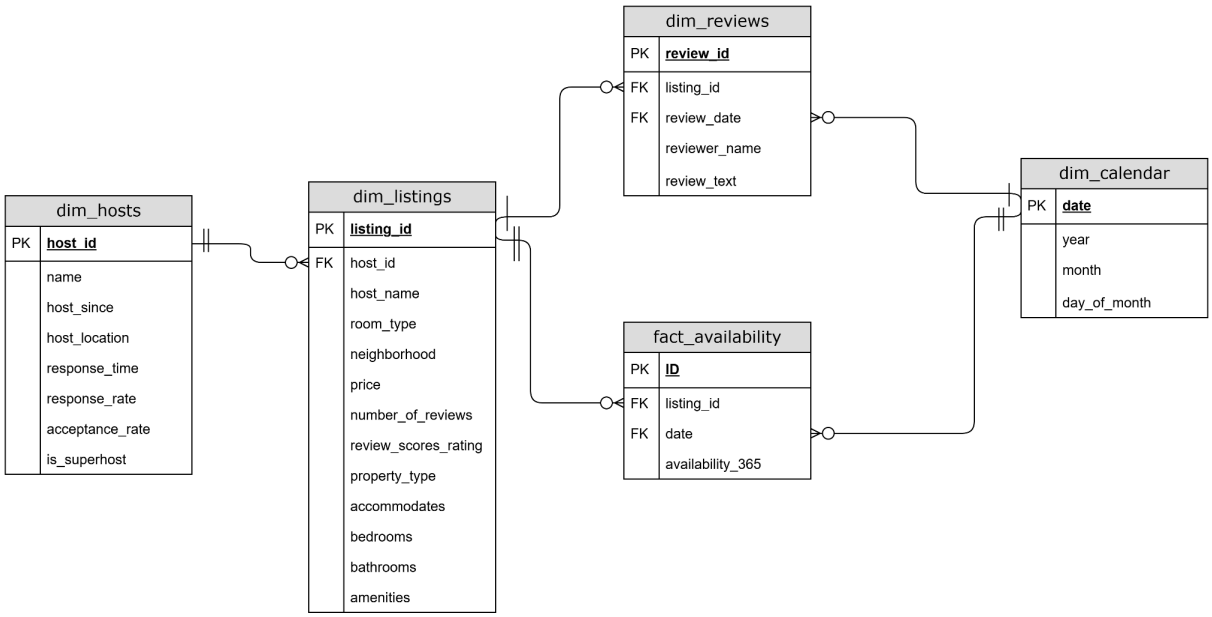


Figure 2.1: Dimensional Model Diagram

Data Dictionary

Fact Table

This table serves as the analytical centerpiece, representing the relationship between individual listings and the dates on which occupancy data is recorded.

Table 3.1: Data Dictionary, Fact Table - Listings

Feature	Description	Type
ID	S/N	Integer
listing_id	Airbnb's unique identifier for the listing	Integer
date	Review date	Date
availability_365	The availability of the listing 365 days in the future as determined by the calendar. Used to determine the occupancy rate	Integer

Dimension Tables

1. dim_calendar - Enables time-based analysis of availability trends.

Table 3.2: Data Dictionary, dim_calendar

Feature	Description	Type
Date	Review date	Date
year	Year	Integer
month	Month	Integer
day_of_month	Day of the month	Integer

2. dim_listings - Provides detailed attributes of each listing

Table 3.3: Data Dictionary, dim_listings

Feature	Description	Type
listing_id	Airbnb's unique identifier for the listing	Integer
host_id	Airbnb's unique identifier for the host/user	Integer
host_name	Name of the host	String
room_type	Entire place/private room/shared room	String
neighborhood	Host's description of the neighbourhood	String
price	Daily price in local currency	Float
number_of_reviews	The number of reviews the listing has	Integer
review_scores_rating	Review scores for the listing	Float
property_type	Self selected property type	String
accommodates	The maximum capacity of the listing	Integer
bedrooms	The number of bedrooms	Integer
bathrooms	The number of bathrooms in the listing	Integer
amenities	Amenities available in the listing	String

3. dim_hosts - Captures profile information for hosts.

Table 3.4: Data Dictionary, dim_hosts

Feature	Description	Type
host_id	Airbnb's unique identifier for the host/user	Integer
name	Name of the listing	String
host_since	The date the host/user was created	Date
host_location	The host's self reported location	String
response_time	That duration at which a host respond to enquiries	String
response_rate	That rate at which a host respond to enquiries	String
acceptance_rate	That rate at which a host accepts booking requests	Integer
is_superhost	Host has Superhost status in Airbnb	Boolean

4. dim_reviews - Holds guest feedback for each listing.

Table 3.5: Data Dictionary, dim_reviews

Feature	Description	Type
review_id	Unique identifier for the review	Integer
listing_id	Airbnb's unique identifier for the listing	Integer
review_date	Review date	Date
reviewer_name	Name of reviewer	String
review_text	Content of review	String

Data Profiling

The data profiling stage is focused on acquiring an initial understanding of the dataset. This includes an assessment of the data schema, distribution of values, and overall structure. Key activities in this phase include:

- Identifying variable types and formats
- Detecting anomalies, such as outliers or inconsistencies
- Assessing data completeness and uniqueness
- Evaluating the relevance of each feature for the intended analysis

This foundational step ensures familiarity with the dataset and informs decisions in subsequent cleaning and transformation phases.

Temporal Coverage

The ‘host_since’ feature ranges from 2009-06-29 to 2024-09-14, indicating that the data includes hosts who have been active on Airbnb for over 15 years. Similarly, the ‘last_review’ feature ranges from 2014-06-28 to 2024-09-26, providing over 15 years of review data. This extensive time span offers a strong foundation for analyzing historical trends in host behaviour and guest feedback.

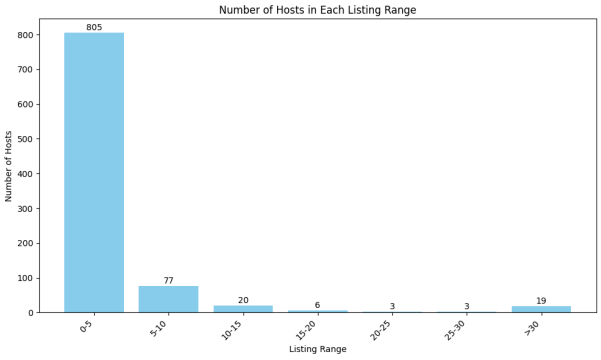


Figure 3.1: Number of Host in Each Listing Range

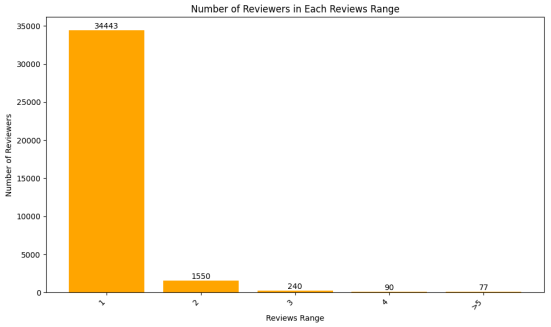


Figure 3.2: Number of Reviewers in Each Review Range

As shown in Figure 3.1, the majority of the hosts manage between one to five listings, indicating a diverse but predominantly small-scale host population similar to SingStay. Additionally, Figure 3.2 shows that most reviewers have submitted only a single review. This suggests a lower possibility of review manipulation or bias from repeat reviewers or from the hosts themselves.

Geographic Distribution

The dataset includes listings from all regions of Singapore, with the following distribution shown in Figure 3.3 below. The Central Region accounts for the majority of the listings, which aligns with expectations as it is the core of Singapore's business, commercial, and tourism activities. This geographic concentration is likely to influence occupancy patterns and pricing strategies, making it a critical consideration in model interpretation.

REGION	NUMBER OF LISTINGS
Central Region	2732
West Region	248
East Region	205
North-East Region	117
North Region	79

Figure 3.3: Number of Listings in Each Region in Singapore

Text Data

The dataset used contains several features that contain text data. These data will be used later on for the text analytics portions.

Features in Listings Data

1. **Listing Name**: Title of listing as displayed on airbnb website

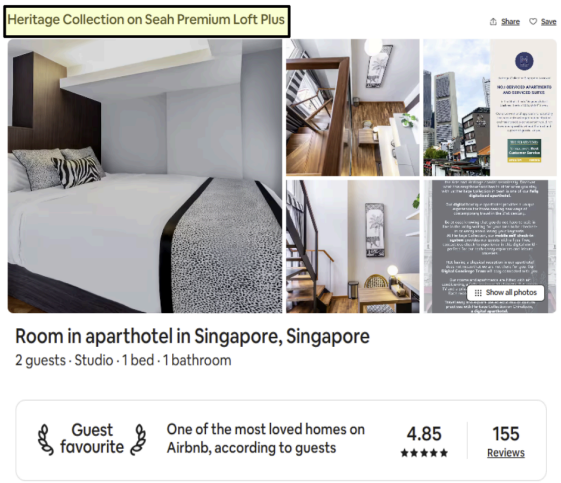


Figure 3.4: AirBnb Screengrab - Title of Listing

2. **Listing Description**: Description of listing provided by listing owner.

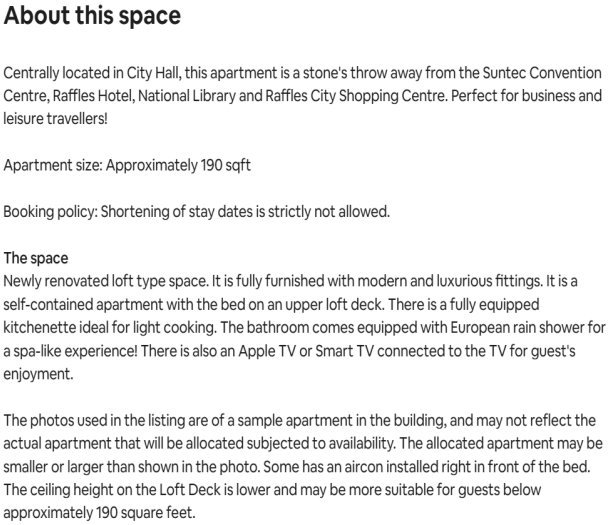


Figure 3.5: AirBnb Screengrab - About this Space

3. **Neighbourhood Overview Description**: Description of neighbourhood listing is situated as per listing owner.

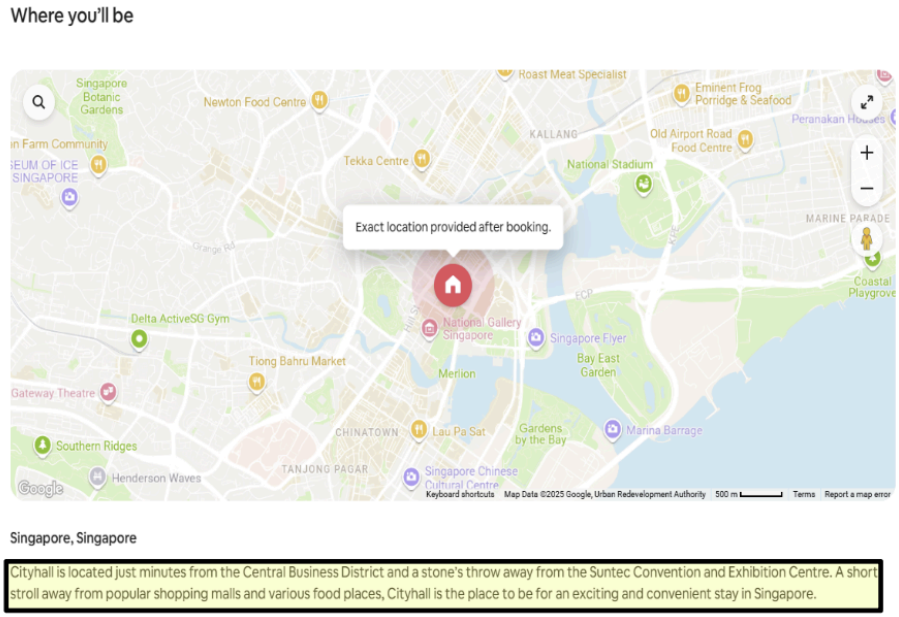


Figure 3.6: AirBnb Screengrab - Neighbourhood Overview Description

4. **Host About**: Hosts About Section describing the host's personality or hosting style, as written by the host.

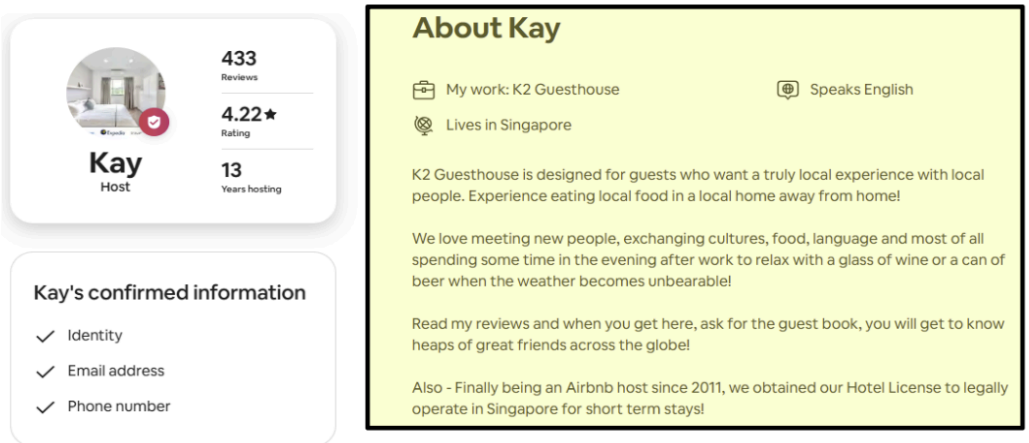


Figure 3.7: AirBnb Screengrab - Host About Section

Features in Reviews Data

1. **Comments**: Feedback from guests about their stay at the listing.

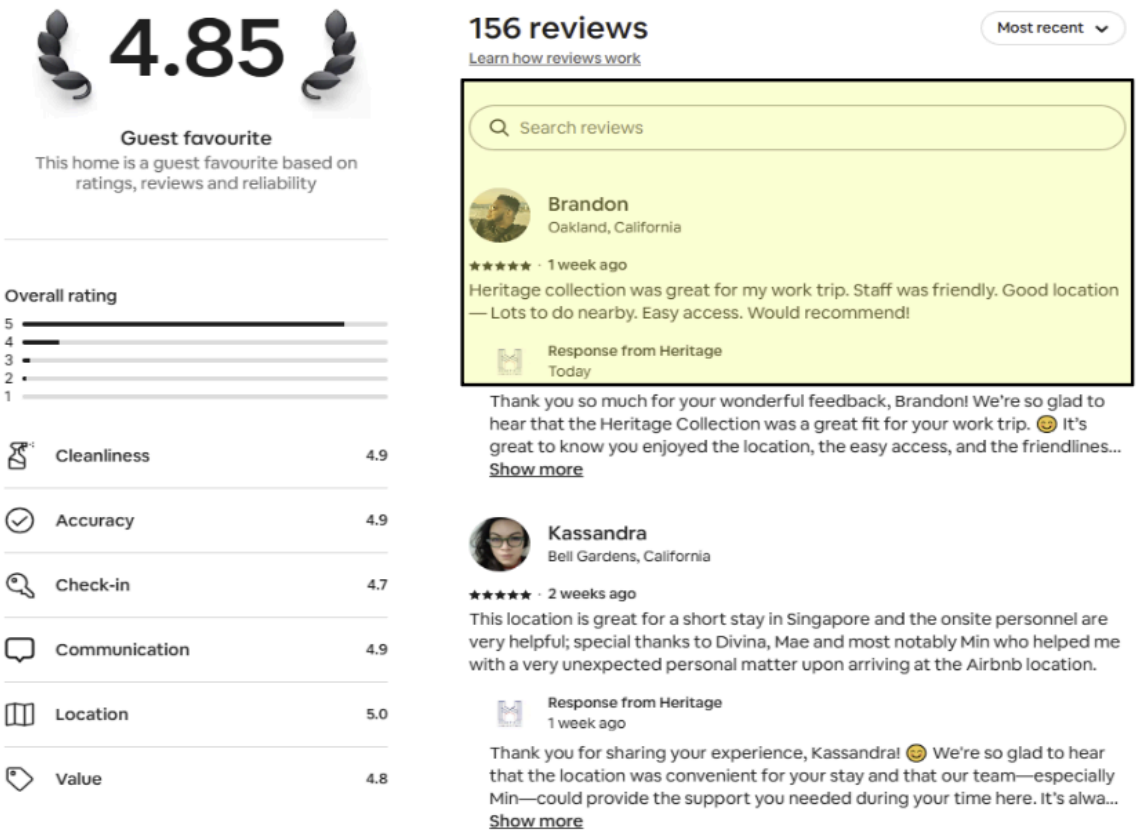


Figure 3.8: AirBnb Screengrab - Comments from Guests

Methodology

This section outlines the methodology adopted for the project, including the Exploratory Data Analysis (EDA), text analysis, and building of the text models and predictive models.

Exploratory Data Analysis

The objective of EDA is to gain a comprehensive understanding of the data's structure, quality, and underlying patterns, thereby enabling informed decisions in subsequent modeling and analysis stages. The EDA process in this context consists of three main phases: Data Cleaning, Data Pre-Processing, and Data Exploration.

Data Cleaning

Data cleaning is a critical phase aimed at improving the quality and usability of the dataset. These steps are essential to ensure the accuracy and reliability of downstream analyses. The procedures were conducted as follows:

Dropping Irrelevant Columns

Features that offered no analytical value or were redundant were excluded. This helps to streamline the data and reduce noise that may impact model performance.

Data Type Conversion

Several variables were cast to appropriate data types (e.g. integers, floats, dates) to ensure consistency and to facilitate correct downstream operations such as modeling and visualisation.

Handling Missing Values

Based off the context, missing or null values across both numerical and categorical features are managed as follows:

- **Replacing blanks with “unknown”**:
  For attributes such as ‘host_location’ and ‘bedrooms’, blank entries were retained and labeled as "unknown" rather than removed. This decision was made to preserve listings that may lack complete information but are still valid.

- **Value extraction from related columns**:
  For example, both ‘bathrooms’ and ‘bathrooms_text’ contain overlapping information, but many entries had missing values in one or the other. They were combined into a single bathroom count feature.

- **Dummy values from missing dates**:
  In features such as ‘first_review’ and ‘last_review’, missing entries were filled with a placeholder date, 1999-01-01. This allowed us to retain these records for analysis while allowing these data to be filtered off when necessary.

Error Removal

Logical inconsistencies and format-based errors (e.g. symbols in numeric features) were identified and corrected to ensure data integrity.

Data Pre-Processing

Following cleaning, the data was pre-processed and prepared for analytical procedures. This phase enhances the data's compatibility with analytical tools and modeling techniques. Pre-processing activities included:

Text Translation

Text features were translated where necessary to ensure language consistency across records. “googletrans==4.0.0=rel” API is used to detect and translate.

Encoding

Categorical variables were encoded into numerical formats suitable for machine learning algorithms. Two encoding techniques were applied:

- **Label Encoding**:
  Features such as ‘neighbourhood_cleansed’ were encoded using scikit-learn’s “LabelEncoder”, which assigns unique integer values to each category. The encoder was saved using “pickle” to ensure consistency during future model retraining or deployment.
- **Manual Dummy Encoding**:
  For categorical features that have few categories, for instance ‘host_is_superhost’ with ‘t’, ‘f’, ‘uk’, they were manually converted to numerical “[1, 0, 2]” respectively. These allow for easier conversion and interpretation.
- **One-Hot Encoding**:
  High-cardinality features like ‘amenities’ were one-hot encoded. The top 50 most frequent amenities were retained to reduce dimensionality. A binary matrix was then generated to indicate the presence or absence of these selected amenities in each listing.

A specialized one-hot encoding with aggregation was also applied to the ‘RTopic’ feature (from topic modeling) to handle multiple reviews per listing. This transformation is described in detail in the Predictive Modeling section.

Data Transformation

Relevant transformations were applied to normalize values, derive new variables, and structure data for improved interpretability.

Data Exploration

In the final stage of EDA, the dataset was explored in depth to uncover trends, patterns, and preliminary insights. This included the following:

Data Validation

Ensuring the accuracy and logical coherence of the data based on defined rules and expectations.

Correlation Analysis

To access multicollinearity among numerical features, we conducted a correlation analysis using Pearson correlation coefficient. A threshold of 0.7 was used to identify pairs of highly correlated variables. The analysis revealed strong correlations in the following feature groups:

- Minimum and maximum night-related variables
- Review score-related metrics
- Availability indicators
- Host-related statistics

More detailed analysis is available below in the predictive model, feature selection section. A visual heat map illustrating the identified correlations can be found in Appendix B.

Text Analytics

Text analytics techniques were used to derive meaningful insights from unstructured textual data. Text analytics was applied to analyse Airbnb listing descriptions and guest reviews. Several text analytics techniques were undertaken such as word clouds, top-tokens analysis and n-gram analysis.

Text Data Tokenisation

To prepare the data for analysis, all the preprocessed text features were tokenised. Tokenisation involves breaking down the text into individual words or 'tokens'. This step is fundamental in natural language processing as it enables the measurement of word frequency and pattern detection.

For example, a listing description like “Spacious modern apartment near MRT” would be split into tokens: ["spacious", "modern", "apartment", "near", "MRT"].

Tokenisation also helped in normalising the text by removing unnecessary punctuation or empty entries and standardising the format across listings and reviews.

Word Clouds: Listing and Review Features

To visualise the prominent words used across listing and review texts, word clouds were generated. These visual representations illustrate the frequency of words through size: the larger the word appears, the more often it is used. Word clouds were created for each of the four listing-related features and the reviews feature (comment_tokens) which gives a quick visual summary of dominant language themes. The following two figures show the visualisation of the tokens in each set of tokens from the respective text features.

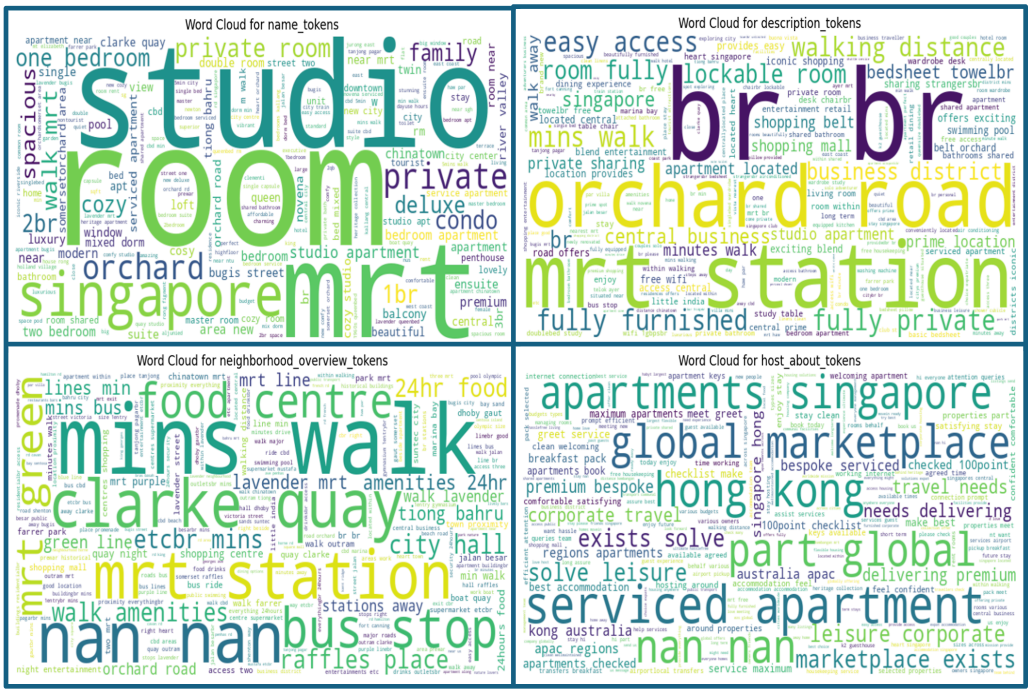


Figure 4.1: Word Cloud - Listings features

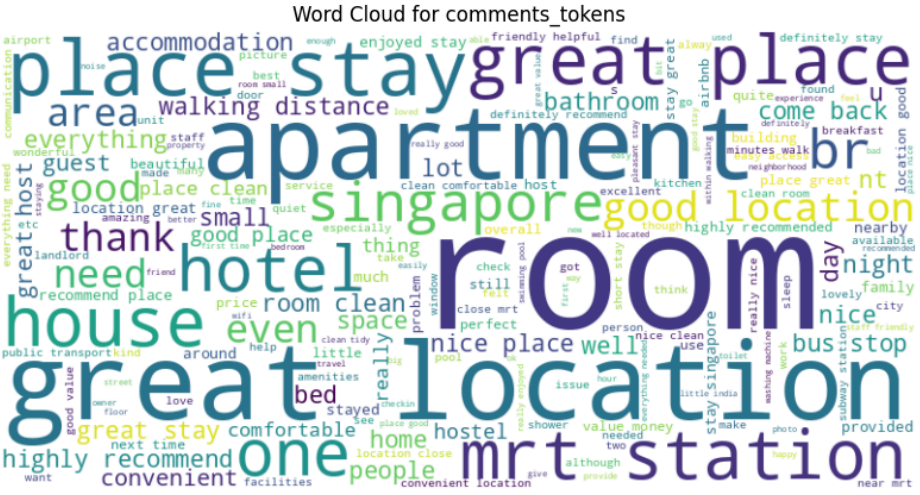


Figure 4.2: Word Cloud - Comment feature

Top 30 Tokens: Listings Features

We identified the most common words used in each of the listing text features. These top tokens highlight the language frequently used by hosts in high-visibility areas of their listings and reflect how all hosts choose to position their properties.

In the ‘name_tokens’ features, the most frequent words include "apartment", "studio", "MRT", and "bedroom", which suggests that hosts place emphasis on the type of accommodation and its proximity to public transport.

In the more detailed ‘description_tokens’ features, common words such as "BR" (Bed Room), "located", "room", "fully", and "stay" appear regularly. This indicates a focus on describing the space layout and the readiness of the property for guests.

For the ‘neighborhood_overview_tokens’, words like "MRT", "walk", "minutes", "city", "food", and "shopping" dominate, demonstrating that hosts often promote convenience and accessibility to amenities as key selling points.

Lastly, in the ‘host_about_tokens’, the most frequently used words include "apartments", "stay", "properties", "Singapore", and "accommodation", revealing a focus on service offerings, hospitality, and geographic branding.

These top tokens not only shed light on common vocabulary used by hosts, however as this analysis is done on all the tokens in each of the text features, they do not offer much actionable insights. In the following section bigram/trigram analysis and correlation of the word analysis to the listings ratings will be explored.

Bigram/Trigram Analysis: Listing Features

Beyond single words, we explored common word pairings (bigrams) and triplets (trigrams) to uncover frequently used phrases. The analysis highlighted distinctive language patterns used by hosts to describe their listings.

In the **‘name_tokens’** feature, frequent bigrams included "bedroom apartment", "studio apartment", and "walk MRT". Trigrams such as "2 bedroom apartment", "5 m walk", and "clarke quay iconic" were common. These combinations reflect how hosts structure listing titles to promote layout and location.

In the **‘description_tokens’**, the top bigrams were more detailed, including "fully furnished", "easy access", and "private sharing". Trigrams like "lockable room fully", "room fully private", and "central business district" emerged, reinforcing descriptions around privacy, location, and business-readiness.

For the '**neighborhood_overview_tokens',** phrases related to proximity and transport were dominant. Bigrams such as "mins walk", "MRT station", and "orchard road" were common, while trigrams included "walk amenities 24hr", "mrt green line", and "mins walk lavender". These expressions clearly indicate that neighbourhood convenience is heavily marketed.

The '**host_about_tokens'** contained a notable number of marketing-style repeated phrases. Nearly all top bigrams and trigrams stemmed from templated service descriptions, such as "serviced apartments singapore", "global marketplace exists", and "bespoke serviced apartments", indicating that hosts (particularly companies) use standardised branding language across listings.

This analysis of n-grams provides a richer understanding of the communication strategies used by hosts. Themes such as location proximity, modern furnishing, and business-readiness were consistently surfaced, suggesting that listings which effectively combine these themes in natural language may be more attractive to prospective guests.

Threshold Score for High/Low Rated Listings

To further examine how textual features relate to performance, listings were categorised based on their review scores. We applied a threshold of 4.8 out of 5 to differentiate between high-rated and lower-rated listings. This decision was grounded in Hospitable Team's: Airbnb Ratings A Host's Guide article (Hospitable Team, n.d.). A 5-star rating means the experience met or exceeded expectations, while a 4-star rating though still positive implies some level of disappointment. A 3-star or below signals a clearly negative experience. A threshold of 4.8 represents the upper tier of guest satisfaction and aligns with hospitality best practices where even small rating differences impact listing visibility and booking rates.

This threshold was further supported by distribution plots of review scores, which revealed a skew toward the higher end of the scale common in platforms where satisfied users are more likely to leave reviews. A cutoff at 4.8 captures truly exceptional listings while allowing for meaningful comparative analysis.

TF-IDF

Before comparing the bigram and trigram patterns between high- and low-rated listings, we applied Term Frequency-Inverse Document Frequency (TF-IDF) weighting to the tokenised text columns. TF-IDF adjusts raw word counts by considering how common or rare a token is across all listings, amplifying the influence of distinctive phrases while down-weighting generic words that appear everywhere. This step was crucial because raw frequency alone could overemphasise common but less informative terms like "apartment" or "stay".

By precomputing TF-IDF scores separately for each textual feature, we ensured that the importance of each n-gram was based not just on how often it appeared, but how uniquely it characterised a listing. The weighted n-grams thus reflect the truly distinctive language patterns that differentiate high-performing listings from lower-performing ones. This foundation enabled a more meaningful and interpretable comparison of bigram and trigram usage across rating categories.

Bigram/Trigram Correlation to Rating: Listing Features

When comparing bigrams and trigrams across high- and low-rated listings, several patterns emerged:

- **Name Tokens**:
  High-rated listings often featured compact, appealing names like “one bedroom”, “walk mrt”, and “cozy studio”, highlighting convenience and comfort. Low-rated listings were dominated by generic or repetitive terms like “private room”, “bedroom apartment”, and “studio apartment”, suggesting less distinctive branding and weaker first impressions.

- **Description Tokens**:
  High-rated listings often highlighted privacy and security through phrases like “fully private”, “room fully”, and “lockable room”, reinforcing guest comfort and independence. Trigrams such as “room fully private” and “lockable room fully” further stressed exclusivity.
  In contrast, low-rated listings emphasised convenience with generic terms like “easy access”, “mrt station”, and “apartment located”, focusing on location rather than personal comfort. This suggests a less distinctive and guest-focused experience.

- **Neighbourhood Overview Tokens**:
  High-rated listings emphasised walkability and local access with phrases like “mins walk”, “walk lavender”, and “food centres”, creating a sense of convenience and authentic neighbourhood experience. Low-rated listings also referenced proximity, but often relied on broader terms like “mrt station” and “central business”, offering less personalised and distinctive local context.

- **Host About Tokens**:
  High-rated listings used phrases like “rent rooms behalf” and “various owners”, giving a more personal, home-like feel to the hosting. Low-rated listings leaned heavily on corporate-sounding language such as “serviced apartments singapore” and “corporate travel needs”, suggesting a more commercialised, less individualised experience.

- **Overall Insight**:
  High-rated listings consistently used language that was specific, inviting, and guest-focused, highlighting privacy, comfort, walkability, and personal hosting. In contrast, low-rated listings often relied on broad, generic descriptors related to location or amenities, lacking distinctiveness and emotional appeal. This divergence in language suggests that high-performing hosts crafted more personalised and memorable experiences, which likely contributed to higher guest satisfaction and ratings.

Top 30 Tokens: Review Features

Next for the review tokens, we analysed the most frequently used words in the guest review comments. These review tokens provide valuable insight into what guests remember and choose to highlight after their stays, offering a candid view of what truly matters to them.

Among the most frequently mentioned words were "place", "stay", "great", and "room", indicating that guests commonly describe the overall experience and space. Words such as "location", "good", "clean", and "nice" also featured prominently, suggesting that hygiene, geographical convenience, and a generally pleasant environment are key contributors to guest satisfaction.

Additionally, terms like "host", "helpful", "friendly", and "recommend" highlight the critical role that host interaction and service quality play in shaping a positive guest experience. The frequent appearance of words like "MRT", "station", "walk", and "convenient" reinforce the idea that proximity to transport and ease of access are consistently appreciated.

Overall, these tokens paint a picture of what guests value most: a clean, comfortable space, a supportive and responsive host, and a location that is easy to reach and well-connected. This reinforces the need for hosts to focus on not only the physical features of the listing but also their communication and service style, all of which leave lasting impressions on guests.

Bigram/Trigram Correlation to Rating: Review Feature

We applied TF-IDF weighting to guest review comments to identify the most distinctive bigram and trigram phrases across all listings. Unlike the listing descriptions, reviews were not separated into high- and low-rated groups for this analysis. Instead, we focused on uncovering common linguistic patterns that shape guest perceptions overall.

The top-ranked bigrams, such as “great location”, “place stay”, and “great place”, reflect standard expectations for a positive stay, with location emerging as a consistent theme. Similarly, leading trigrams like “great place stay” and “good place stay” emphasise satisfaction with both the property and the experience.

Beyond these baseline phrases, expressions like “highly recommend place”, “really enjoyed stay”, and “would definitely stay” captured deeper emotional engagement, signalling that convenience alone is not enough. Guests value experiences that are memorable and recommendation-worthy. The frequent appearance of words tied to comfort ("clean place", "nice clean") and value ("good value money") also suggests that affordability and cleanliness are important dimensions of guest satisfaction.

Overall, the review language demonstrates that while location remains foundational, emotional resonance, comfort, and perceived value significantly contribute to guests' willingness to praise and recommend a listing.

Topic Modeling

In this section we will describe the topic modeling process to undercover themes within the textual features of the Airbnb listings. By identifying dominant topics, we aimed to better understand the types of competitor listings (through Listing text data) and guest experiences (through Reviews text data) associated with both high-performing and lower-performing properties. Topic modeling enables the extraction of interpretable patterns from large text corpora, providing a structured summary of the types of language and concepts most commonly associated with listing success.

Listing Data Topic Modeling (Using LDA)

We began with topic modeling by first categorising the listings based on their average review scores. Just as in earlier sections, a threshold of 4.8 was applied: listings with a rating of 4.8 or above were labelled as "high" performing, and those below 4.8 were labelled as "low" performing. This threshold aligns with hospitality standards where even minor differences in ratings can significantly impact listing visibility and booking rates, and was supported by a right-skewed distribution observed in review scores.

The text features that are stored as tokens in features ‘name_tokens’, ‘description_tokens’, ‘neighborhood_overview_tokens’ in dim_listings and ‘host_about_tokens’ feature from dim_hosts were combined into a single ‘combined_tokens’ features for each listing in a pandas data frame.

After token combination, a dictionary of unique tokens was created using the Gensim library. Tokens that appeared too infrequently (in fewer than two documents) or too frequently (in more than 95% of documents) were filtered out to enhance the robustness of the model by focusing on meaningful terms.

The Latent Dirichlet Allocation (LDA) model was then trained on the corpus using 10 topics, with 10 passes through the dataset to achieve more stable results. A fixed random state was applied to ensure reproducibility. The output from the LDA model allowed us to infer and interpret the dominant themes that differentiate high-rated from lower-rated listings, providing deeper insight into the types of language and listing attributes that correlate with guest satisfaction.

Listing Data Topic Modeling Results

To interpret the outputs of the LDA model, we used pyLDAvis to visualise the topic space, with a relevance parameter (λ) set to 0.4. This value strikes a balance between showing frequent and exclusive terms, allowing us to interpret each topic meaningfully.

The LDA model identified 10 distinct topics based on the combined textual tokens from each listing. Topics were manually assigned descriptive labels based on the Top-30 most relevant terms per topic. For example, Topic 1 assigned: Budget & Student-Friendly Stays. This topic was characterised by terms such as university, budget, professionals, cosy, and quiet, indicating listings that cater to students and young travellers looking for affordable and peaceful accommodations. Table 5.1 below summarises the topics and the labels our team assigned to them.

Table 5.1: Listings Topic modeling Results

Topic ID	Topic Label / Name Assigned	Words Associated with Topic
Topic 0	Luxury Travel in Prime Districts	orchard, iconic, bespoke, marketplace, prime, belt, experiences
Topic 1	Budget & Student-Friendly Stays	university, professionals, budget, quiet, young, peaceful, cosy, grad, stores, awaybr, name
Topic 2	Cultural Food & Local Exploration	geylang, kallang, chinatown, delicacies, stalls, exploring, inn, pub, sheng siong
Topic 3	Heritage Neighbourhoods & Public Transit	lavender, jalan, historical, bus, gymnasium, mustafa, bedsheet, walk, prewar, green
Topic 4	CBD Access & Short-Term City Stays	outram, shenton, clarke, pagarbr, queensize, exit, patio, mins, walk, buildingbr, 24hours
Topic 5	Expat-Ready & Furnished Rentals	expats, ezspace, tiong, novena, bras basah, national, cityhall, licensed, nights, housekeeping, bencoolen, solutions
Topic 6	Leisure & Boutique Lifestyle Stays	holland, village, sentosa, resort, indie, artisanal, gastronomical, wellmaintained, met, quincy, treasures, quaint
Topic 7	Nature-Focused Stays Near Universities	kent, ridge, nature, buona vista, nus, insead, hort, par, heritagebr, beer, vista, walks, local
Topic 8	Flexible Co-living for Nomads	communal, shared, movein, hasslefree, globally, cities, furnishings, company, mattress, stove
Topic 9	Boutique Hostels & Instagrammable Stays	bugis, hostel, capsule, eclectic, igworthy, spiral, civic, harmonising, meadows, picture

To assign each listing a dominant topic, we first calculated the topic distribution per listing using its combined tokens. The dominant topic was then extracted as the one with the highest probability weight. We then grouped listings by rating category ("high" vs "low") and their corresponding dominant topics, enabling us to analyse how topics were distributed across performance levels.

We plotted a bar chart shown in Figure 5.1 that revealed that Topic 5 (Expat-Ready & Furnished Rentals), Topic 6 (Leisure & Boutique Lifestyle Stays), and Topic 9 (Boutique Hostels & Instagrammable Stays) were the most common among high-rated listings. This suggests that well-furnished, visually appealing, and lifestyle-oriented listings tend to be more positively reviewed. In contrast, Topic 5 and Topic 6 also had high frequencies in low-rated listings, indicating that while these themes are popular, execution quality and guest expectations likely differentiate outcomes. The chart generated using seaborn below provides a clear comparison of topic dominance between high and low-rated listings.

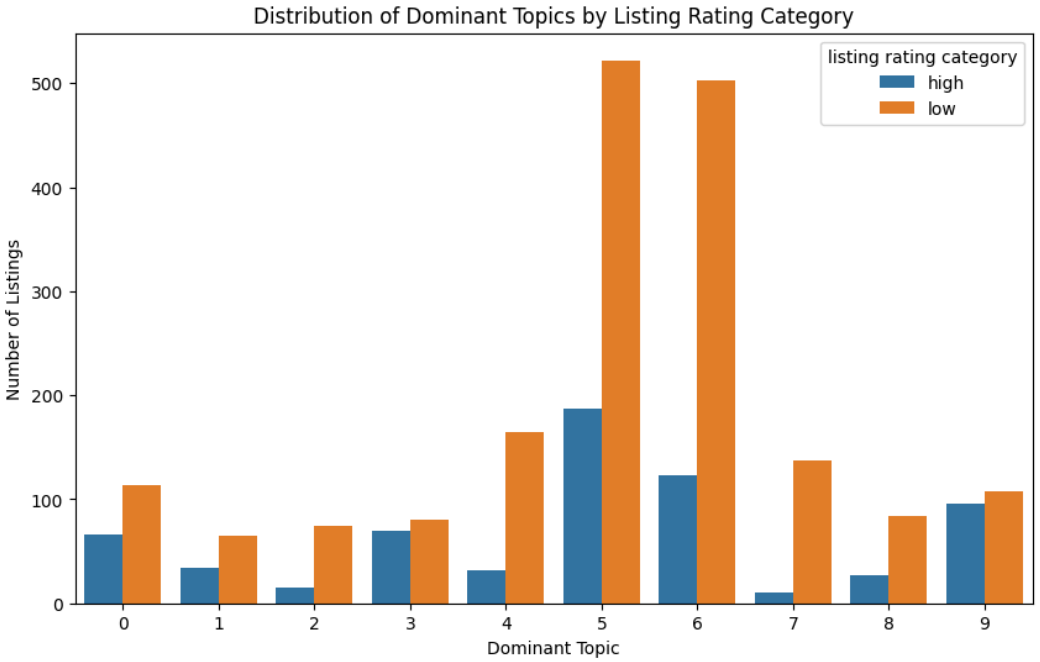


Figure 5.1: Distribution of Listings Across Dominant Topics by High/Low Listing

Review Data Topic Modeling (Using LDA)

A separate LDA model was trained on tokenised guest review comments to uncover recurring themes in guest experiences. Using 10 topics and the same modeling parameters as the listing analysis, the process involved filtering valid review tokens, constructing a dictionary, and generating a bag-of-words corpus. This allowed us to extract common patterns in guest narratives, offering direct insights into what guests consistently mention.

Review Data Topic Modeling Results

Similarly, we used pyLDAvis to identify the top-30 most relevant terms for each topic assigned at lamda = 0.4 and assigned topic labels to the topics. Table 5.2 shows the topic label assigned to each topic.

Table 5.2: Listings Topic modeling Results

Topic ID	Topic Label / Name	Words Associated with Topic
Topic 0	Host Interaction & Overall Experience	us, time, thank, stay, host, family, comfortable, recommend, best, help, always
Topic 1	Location & Transport Convenience	close, mrt, subway, location, walk, near, distance, stop, metro, restaurants
Topic 2	Positive Highlights – Perfect Stay	great, location, recommend, excellent, perfect, amazing, friendly, well, comfortable, responsive, convenient
Topic 3	Unpleasant / Negative Room & Sleep-Related Issues	noisy, sleep, bathroom, noise, air, small, first, bad
Topic 4	Check-in & Hotel-Style Setup	hotel, desk, pictures, key, bed, service
Topic 5	Compact, Clean, and Quiet Rooms	clean, nice, quiet, capsule, tidy, comfortable, bathroom
Topic 6	Facilities & Amenities	pool, kitchen, washing, machine, laundry, dryer, equipped, tv, available
Topic 7	Value for Money	value, price, money, accommodation, excellent, cost, effective, hostel, service, worth, reasonable, maintenance, backpackers, inn
Topic 8	Smooth Process & Communication	checkin, landlord, clear, communication, process, instructions, owner, self, flexible, smooth, hygienic, responded,
Topic 9	Unique & Personal Touches (Quirky Reviews)	villa, landlords, homestay, butler, attention, strange, distinctive, feedback, man, boss, couples, container, units, lorraine

Similarly each review was assigned a topic distribution using its bag-of-words representation, and the dominant topic was extracted based on the highest probability. This enabled us to identify which themes were most frequently mentioned across the review corpus. A bar plot of topic frequencies (Figure 5.2) was generated using seaborn to visualise the distribution of dominant topics, offering a concise summary of recurring guest sentiments and focal points in post-stay feedback.

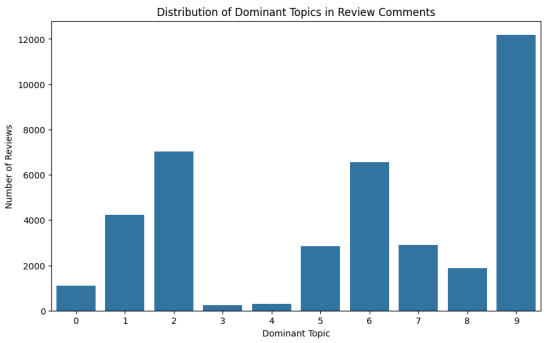


Figure 5.2: Distribution of Reviews across Dominant Topics

Sentiment Analysis

Two sentiment analysis methods are investigated in this project to extract sentiment-related insights from guest feedback. They are TextBlob, a rule-based natural language processing tool,and BERT, a deep learning-based model from Hugging Face's 'nlptown/bert-base-multilingual-uncased-sentiment'. This analysis will allow us to analyse how guests feel, for instance, distinguishing between frequently mentioned features that are praised versus those that are criticized. Incorporating sentiment alongside topic modeling enhances our ability to identify not only recurring themes but also their qualitative impact on guest satisfaction and occupancy performance.

Table 6.1 below shows the comparison of the results from both models to evaluate the model consistency. Both models classified the vast majority of reviews as positive, with a minimal difference in the number of neutral and negative reviews. Overall, 89.73% of sentiment predictions matched across both methods, indicating a high level of agreement.

Table 6.1: Sentiment Classification Comparison Between TextBlob and BERT

Sentiment Category	TextBlob (Count)	BERT (Count)	Difference
Positive	36215	36348	+133
Neutral	1795	1678	-117
Negative	1272	1256	-16
Agreement Rate	**89.73%**		

Since the correlation is strong and that TextBlob polarity and subjectivity scores can be useful for interpretability, we used results from TextBlob for downstream analysis. Its lightweight nature, consistent classification patterns, and ease of scalability across large text corpora made it ideal for extracting broad sentiment trends. Although TextBlob-derived polarity and subjectivity scores were not direct inputs in the final predictive model, this sentiment analysis phase provided significant value. It allows for deeper contextualisation of topics such as comfort, location, and service responsiveness, which are relevant to both user experience and business decisions. Our future recommendations will include utilizing polarity and subjectivity score analysis to inform a sentiment-based monitoring system.

Predictive Model

This section covers the predictive model developed to identify key factors influencing Airbnb occupancy rates.

Target Variable - Occupancy Rate

The target variable for the analysis is occupancy rate, that is not directly available in the dataset. Hence, we have derived it using the feature 'availability_365' using the formula as follows:

$$occupancy\_rate = 1 - (\frac{availability\_365}{365})$$

The 'availability_365' captures long-term booking patterns, providing an annual perspective on listing availability that can be used as a proxy of occupancy. The assumption that fewer available days implies higher occupancy.

Feature Selection

To ensure an efficient and reliable model, we selected features using a combination of statistical analysis and domain knowledge.

For correlation matrix analysis, we computed the Pearson correlation coefficients between numerical features as mentioned in the data exploration section. We then removed variables with a high correlation threshold of above 0.7 to reduce multicollinearity. In each pair of highly correlated variables, we kept the feature with stronger relevance to occupancy prediction based on business logic or interpretability.

We opted not to use backward elimination for correlation-based feature reduction as it may not generalize well to the tree-based models like Random Forest or Gradient Boosting . Our focus is on simplifying the feature set upfront rather than relying on automated elimination that could remove features valuable for interpretation or non-linear relationships. This approach aligns with (Kuhn and Johnson, 2013), who recommend correlation filtering and domain expertise as a more transparent alternative to automated feature selection techniques, particularly in business-focused predictive modeling.

In the application of domain knowledge, we also removed manually features that are redundant and low in interpretability based on this project. It allows us to preserve features that, while not strongly correlated on their own, might still interact meaningfully with others in ensemble models. A list of features considered and eliminated is provided in Appendix C.

We intentionally chose not to perform feature selection within predictive model training functions as our goal of the model is to interpret key influencing features too. By removing features purely based on algorithmic selection could compromise interpretability and hinder our ability to connect findings with specific, actionable business decisions.

Model Training, Evaluation and Selection

There are three models being trained and evaluated in this project to identify the most suitable model to estimate the Airbnb occupancy rates: Linear Regression, Random Forest and Gradient Boosting. Each model is chosen to represent a different modeling approach.

Data Splitting

In this project, we trained our predictive models using the entire cleaned dataset, which spans over 15 years of listing and review data, from 2009 to 2024. The dataset is divided into 80/20 train-test split, where 80% of the data was used for training and 20% was held out for model evaluation. It is decided based off the following factors considered:

- Dataset size, of about 3,300 listings and 39,000 reviews, remains computationally manageable
- Primary objective of uncovering broad, generalisable patterns in drivers of occupancy rates
    - not tailoring model to specific user segment or narrow time frame

Hence, using the full dataset instead of sampling the data enables maximum information retention, which supports more comprehensive model training and enriched interpretation of feature importance. However, we do acknowledge that the dataset's long temporal range introduces potential feature drift, seasonality effects and behavioural changes over time. For instance, impact of the pre-pandemic and post-pandemic periods. To address these, we will cover in our further improvements section of our report, to incorporate temporal segmentation, stratified sampling and comparative model evaluation across different time slices. This approach aligns with the principles of adaptive modeling in time-evolving systems (Gama et al., 2014; Webb & Copsey, 2011), and can enhance both predictive power and business relevance as the Airbnb market continues to evolve.

Model Training and Evaluation

Table 7.1 below shows the model comparison and evaluation (Kuhn, M., Johnson, K., 2013). The two performance metrics used are the R-squared (R^2): indicating the proportion of variance in the occupancy rate, higher the better; and Mean Squared Error (MSE): measuring the average of the squared differences between actual and predicted values, lower indicates more accurate.

Table 7.1: Predictive Model Comparison (Pre-tuning Results)

Criteria	Linear Regression	Random Forest	Gradient Boost
Model Type	Linear, parametric	Ensemble, bagging-based	Ensemble, boosting-based
Interpretability	High - coefficients are transparent	Moderate - feature importance available	Low - complex structure
Handling of Non-linearity	Poor	Good	Very Good
Overfitting Risk	Low	Moderate	Higher if not tuned
R^2 (Test Set)	0.5599	0.7931	0.7276
MSE (Test Set)	0.0913	0.0429	0.0565
Sensitivity to 80/20 Split	High (prone to variance)	Stable due to ensemble averaging	Sensitive to outliers, needs tuning
Business Interpretability	Very high	Balanced between accuracy and insights	Low
Suitability for This Project	Simple baseline	Best fit - balance of accuracy and insights	Good candidate for future tuning

Model Selection

Random Forest is selected as the final model due to its strong balance of predictive accuracy, robustness, and interpretability. Among the three models trained, it achieved the best performance with the highest R^2 of 0.7931 and the lowest MSE of 0.0421, indicating strong fit and generalisation. In addition, Random Forest handles non-linear relationships, outliers and multicollinearity effectively, which fits our complex dataset. The main reason is that it provides feature importance scores, which leads us to our key objective of understanding the key features that drive occupancy.

It is known that Gradient Boosting and XGBoost deliver higher predictive accuracy in many contexts (Chen & Guestrin, 2016). However, they tend to be more computationally intensive and require careful hyperparameter tuning, and are less interpretable generally (Molnar, 2022). It may limit the usability in our project's aim of adding value to the business by using the analysis to guide decision making and communicate with the stakeholders. To achieve the key goal of this project, to derive actionable insights from model results, the feature importance output from Random Forest (Breiman, 2001) offers a clearer and more accessible interpretation, enabling operational recommendations.

Although Out-of-Bag (OOB) scoring is a valid evaluation method supported by Random Forest for internal performance estimation, we chose to use an explicit 80/20 train-test split combined with cross-validation via GridSearchCV covered in the section below. This decision was made to ensure consistency across the models compared for easier performance benchmarking. In addition, with a dedicated test set on a moderate sized dataset, it allows for more transparent performance comparison and hyperparameter tuning. It helps to evaluate generalization on unseen data more directly.

That being said, in future iterations of this project, we recommend experimenting with XGBoost or LightGBM to assess potential performance improvements. This is especially so as more data becomes available or the predictive accuracy becomes priority over interpretability. Also, we acknowledge that incorporating OOB scoring could offer additional insights and can be included in future iterations for further robustness testing.

Hyper Parameter Tuning - GridSearch

The selected model, Random Forest, is further optimised using GridSearchCV to identify the best hyperparameters via cross-validation. The key parameters tuned included the following, shown with the results returned:

- **n_estimators** (number of trees in the forest): 300
- **max_depth** (maximum depth of each tree): not restricted
- **min_samples_split** (minimum number of samples required to split an internal node): 2

It returned a negative MSE of -0.053 during cross-validation. The final trained model has a MSE of 0.0421 and R^2 of 0.7971 on the test set, indicating strong predictive performance.

Top 15 Features Extraction

Using the feature importance scores from the Random Forest Model, we extracted the top 15 most important features (see Figure 7.1) to interpret the model outputs. These features provide insights as to the key drivers that were most predictive of occupancy rate.

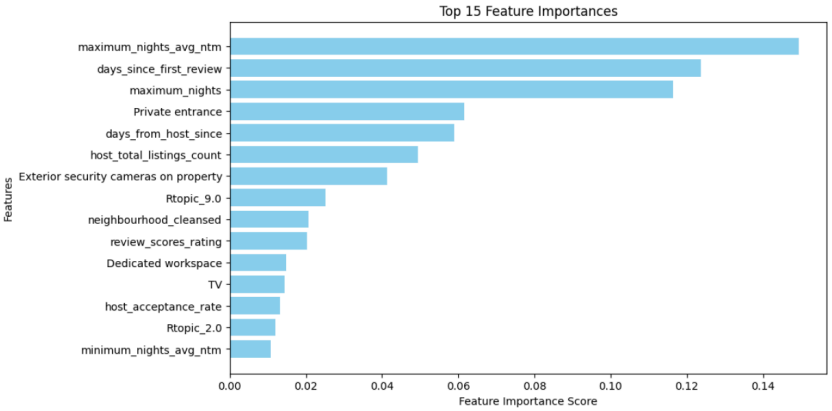


Figure 7.1: Top 15 Feature Importance Most Predictive of Occupancy Rate

Feature importance is used strictly for interpretation and not elimination in this project. We retained all features to avoid discarding variables that may contribute via interaction effects. The analysis will be carried out on these top features to gain insights that can inform actionable business recommendations in the next section, “Key Findings, Insights and Recommendations”.

Key Findings, Insights and Recommendations

With the top 15 features from the optimised Random Forest model (see Figure 7.1), with the inclusion of 'price' feature, an in-depth analysis is carried out to extract actionable insights for improving occupancy below. These insights are grouped by six key categories: amenities, listings statistics, booking policies, review content, location, and pricing. A summary of the key insights and recommendations are tabulated in Appendix D for quick reference and operational use.

These insights are intended to guide strategic decision-making and performance monitoring across SingStay's portfolio of listings. We recommend integrating the analysis and visualisations below into SingStay's internal systems to support ongoing performance tracking. As more listings and reviews accumulate overtime, SingStay can retrain the optimized Random Forest model on updated data to identify the latest top predictors of occupancy.

Although future models may yield slightly different top features depending on guest behaviour or market shifts, the current analysis is still expected to highlight core, consistent drivers of occupancy. This is based on the fact that most features highlighted here are guest-facing attributes such as amenities, host responsiveness and review quality, and fundamental considerations like location and price that are always critical decision points for both guests and hosts. Regardless of changes in seasonality or listing trends, these factors are likely to remain central to guest preference and booking behavior.

Amenities to Include

Several amenity-related features emerged as significant drivers of occupancy. Listings with the following amenities were found to have higher occupancy rates:

- Private entrance
- Exterior security cameras
- Dedicated workspace
- TV

Collectively, it suggests that listings that value safety, privacy and guest comfort catering for work and leisure travelers have better occupancy.

Recommendations on Amenities

SingStay is encouraged to include these amenities on their properties to increase the attractiveness of their listings, improve perceived value and increase booking potential.

Listing Statistics for Monitoring

The feature 'days_since_first_review' appearing in the top 15 features suggests that hosts with longer track record of guest engagement tend to achieve higher occupancy. On top of that, the 'review_scores_rating' and 'host_acceptance_rate' correlate with better performance of their listings, reinforcing the importance of consistent quality and responsiveness in driving bookings.

Recommendations on Listing Statistics

SingStay should build a strong and credible review history and encourage feedback after each stay. Emphasis should be placed on maintaining their service standards, accepting bookings promptly, and responding promptly to inquiries. They can monitor their 'review_scores_rating' and 'host_acceptance_rate' regularly to keep track of and sustain their performance across properties.

Booking Policies

The dataset revealed that booking policies, 'maximum_nights' and 'minimum_nights_avg_ntm', play a role in influencing occupancy. The bar charts in Figure 8.1 below shows the current occupancy rate of the maximum and minimum nights binned. It suggests that overly long maximum night limits may reduce listing attractiveness and that occupancy steadily declines as minimum night requirement increases beyond 60 days. Overall, a moderate flexibility in booking policies (particularly around 8-60 nights) is associated with higher occupancy rates.

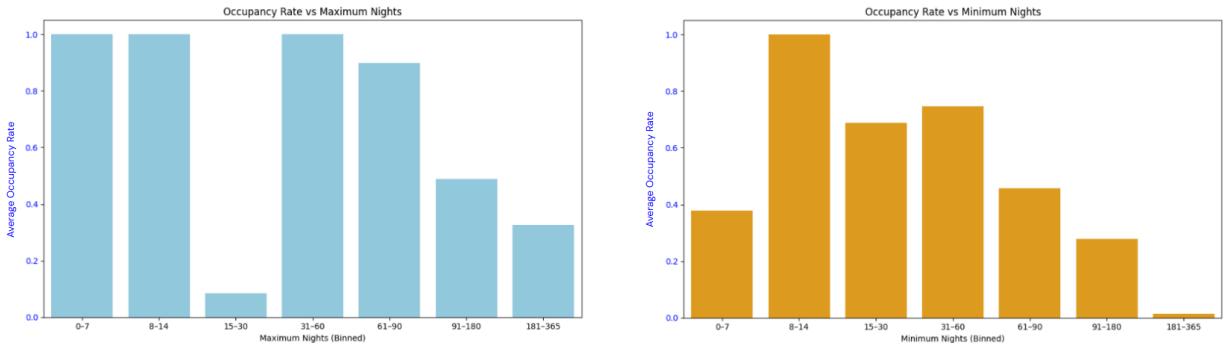


Figure 8.1: Chart of Occupancy Rate vs Maximum and Chart of Occupancy Rate vs Minimum nights

Recommendations on Booking Parameters

SingStay can consider to re-evaluate and experimenting with flexible night stay policies, especially in underperforming listings. A/B testing or periodic adjustments can be implemented to determine optimal configurations across different listing types or regions. SingStay can use this chart to monitor if there are any changes as more data is available.

Review Topics for Consideration

Two review-related themes were associated with higher occupancy from the textual analysis using LDA topic modeling. These topics represent recurring patterns in highly rated reviews and offer actionable insights into what guests value most. The topics with the words group by LDA are as follows:

- RTopic_2: **Perfect Stay**
  This topic highlights universally positive experiences such as great location, comfort and responsiveness.
  - "great, location, recommend, excellent, perfect, amazing, friendly, well, comfortable, responsive, convenient "

- RTopic_9: **Unique and Personal Touches**
  This topic reflects quirky, memorable, and personalized elements that enhance the guest experience.
  - "villa, landlords, homestay, butler, attention, strange, distinctive, feedback, man, boss, couples, container, units, lorraine"

Recommendations on Review Topics

SingStay can enhance guest satisfaction and booking potential by prioritizing hospitality and service standards that align with these themes. Listing descriptions should emphasize location, comfort, and convenience, especially for mainstream units. For properties with the potential for differentiation, SingStay can explore unique thematic offerings (e.g., a Pokémon-themed room, eco-friendly units, or boutique-style decor) to create shareable and memorable stays.

Location

Location plays a critical role in determining a listing's visibility, competitiveness and occupancy performance on Airbnb. A review of regional performance shows clear trends that can inform SingStay's strategic planning and expansion efforts.

Occupancy Rate by Region

As shown in Figure 8.2, listings in the Central Region exhibit the highest average occupancy rates, which is expected given their proximity to tourist attractions, transport hubs, and commercial centers. However, fringe regions such as Marina South and Tuas also show notably high occupancy, despite having fewer listings — indicating promising opportunities for market entry and targeted investment.

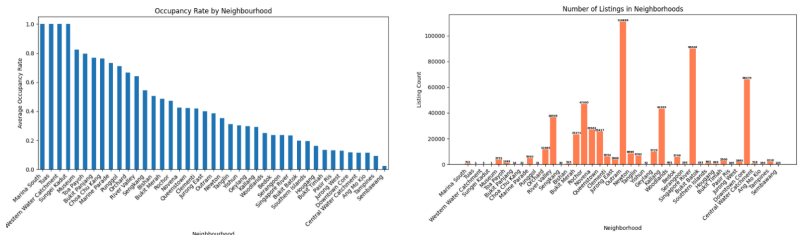


Figure 8.2: Charts of Occupancy Rate by Neighbourhood with Number of Listings in Neighbourhood

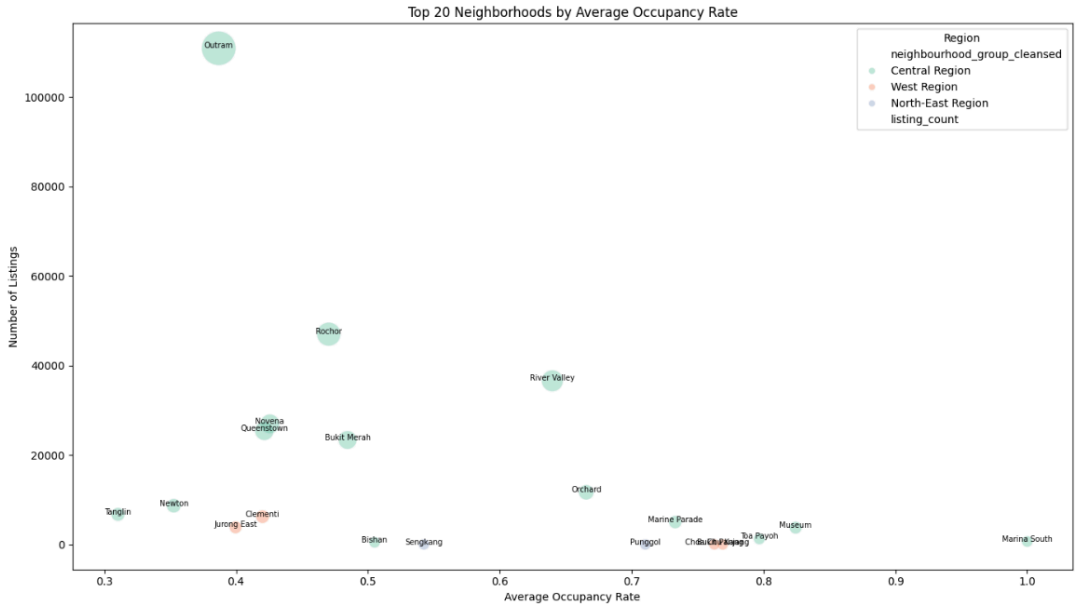


Figure 8.3: Bubble Chart of Top 20 Neighbourhood by Average Occupancy Rate

Figure 8.3 highlights high-performing locations in terms of occupancy and volume of listings. Central areas appear in the Top 20 as expected, but fringe regions like Marina South and Tuas emerge as valuable niches. SingStay can use these visualizations to continuously monitor shifts in occupancy trends as more listings are added and market conditions evolve.

Number of Reviews and Superhost Status

Further supporting analysis (Figure 8.4) indicates that central region listings also receive the highest average number of guest reviews, particularly those with Superhost status. High review volume enhances visibility on Airbnb's platform and builds trust with potential guests.

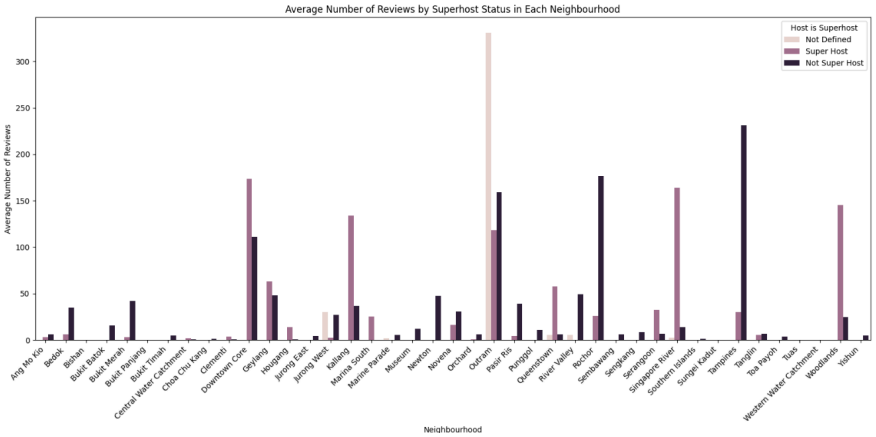


Figure 8.4: Chart of the Average Number of Reviews by Superhost Status in Each Neighbourhood

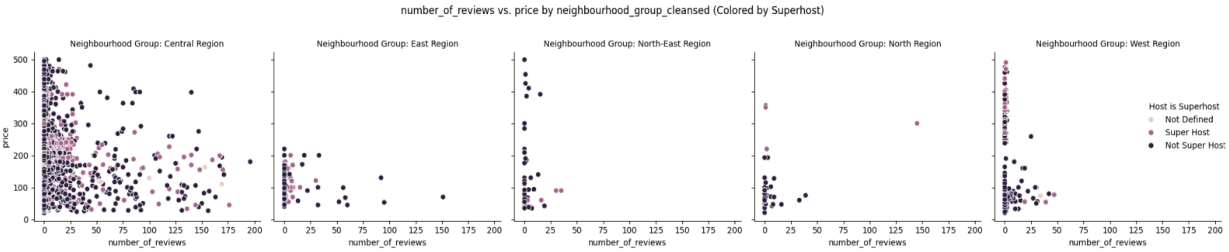


Figure 8.5: Chart of the Average Number of Reviews by Superhost Status in Each Neighbourhood

Additionally, there appears to be a pricing-related trend, where lower-priced listings tend to receive more reviews (Figure 8.5), suggesting a value-conscious guest segment that is more likely to leave feedback. While pricing is an important lever, review generation remains a controllable and strategic factor for improving listing performance across all regions.

Recommendations on Location and Review Engagement

To improve occupancy and market presence, SingStay can consider a multi-pronged strategy focused on regional opportunities and guest engagement. First, the company should explore strategic expansion into high-occupancy fringe areas such as Marina South, Tuas, Western Water Catchment, and Sungei Kadut, where competition remains low but demand indicators are strong. To enhance the attractiveness of these locations, SingStay may partner with tourism agencies such as the Singapore Tourism Board (STB) or collaborate with local transport providers to offer shuttle services to MRT stations or nearby attractions, improving convenience for guests. In parallel, implementing a review follow-up system—such as automated post-stay reminders—can help boost guest feedback, particularly in non-central listings and newly launched properties. Lastly, all listings, especially those outside the Central Region, should strive for Superhost status, which enhances trust, visibility, and competitiveness across Airbnb's platform.

Price

Listings priced between $301–$500 achieved the highest average occupancy, as shown in Figure 8.6. Interestingly, 1-room apartments, although priced higher, consistently maintained strong occupancy—particularly in non-central regions (Figure 8.7). This aligns with word cloud results, which indicated a guest preference for privacy.

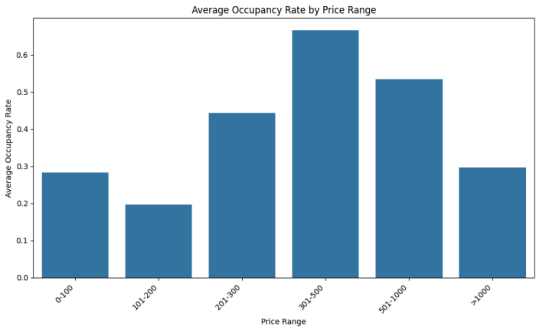


Figure 8.6: Occupancy Rate by Price Range

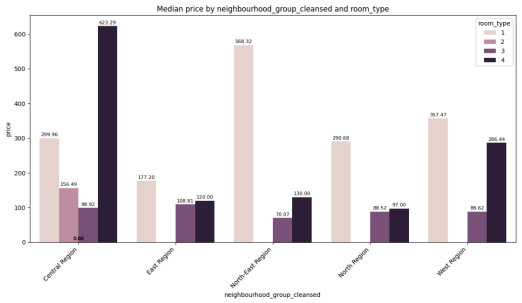


Figure 8.7: Price by Neighbourhood and Room Type

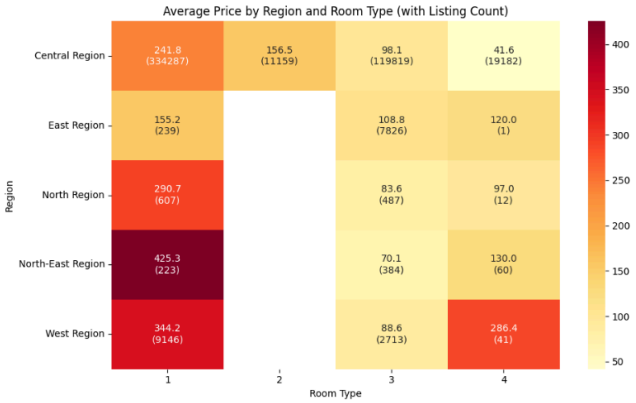


Figure 8.8: Heatmap of Average Price by Region and Room Type

Additionally, the heatmap in Figure 8.8 demonstrates that 1-room listings are consistently the highest-priced room type across all regions. The heatmap data was cleaned using the Interquartile Range (IQR) method, retaining only the central 50% of values to reduce the influence of outliers. While the prices shown should not be treated as absolute pricing recommendations, they provide valuable benchmarks for comparison and strategy refinement.

Recommendations on Pricing

SingStay should consider optimizing listing prices by referencing data-informed price bands, particularly around the $301–$500 range. Additionally, 1-room accommodations should be maintained and promoted across regions, given their strong performance and alignment with guest preferences for privacy and value.

Further Analysis & Future Recommendations

This project establishes a strong foundation for identifying key drivers of Airbnb occupancy. However, there are several opportunities to enhance model precision, adaptability, and robustness in future iterations. In addition, the key limitation faced was the limited data for analysis. As a new entrant into the short-term rental market for only a year, SingStay possesses very small data. However, with continued operation, SingStay will accumulate more listings and reviews data. This enables future analysis using techniques such as temporal segmentation or rolling time windows to analyse more recent trends or seasonal occupancy fluctuations. While the following approaches were not applied in the current phase, primarily due to project scope, dataset size, and the emphasis on model interpretability, they offer valuable directions for continued development.

Temporal Segmentation and Trend Analysis

To account for seasonal trends and temporal variability, future models can incorporate rolling time windows or segmented time periods (e.g., pre-/post-pandemic, recent 3 years vs. full dataset). This will enable more time-relevant insights and help assess how feature importance evolves over time.

Stratified Sampling and Balanced Data Representation

As the current model was trained on the full dataset, stratified sampling can be introduced in future work to ensure balanced representation across listing types, price bands, regions, and time periods. This will help improve model generalization and reduce bias in underrepresented groups.

Comparative Model Evaluation by Time Slice

Running models across different temporal subsets (e.g., recent data only vs. entire dataset) can reveal how time-based variability affects model performance and feature relevance, improving both accuracy and interpretability.

Algorithm Enhancement and Benchmarking

Although Random Forest offers a strong balance of accuracy and interpretability, future iterations may explore more advanced algorithms like XGBoost or LightGBM, which are known for their superior predictive performance. These can be tested as the dataset expands, particularly if prediction accuracy becomes a higher priority than interpretability.

Incorporating Out-of-Bag (OOB) Evaluation

While this project used an 80/20 train-test split with cross-validation for consistency across models, future work could incorporate OOB scoring, particularly for ensemble models like Random Forest. This would offer an additional layer of performance validation and help verify model robustness without sacrificing test set integrity.

Sentiment-Based Monitoring System

In future iterations, SingStay can implement a sentiment-based feedback dashboard to track guest sentiment trends over time at the property or host level. This can support early detection of service quality issues, inform listing description updates, and guide targeted service improvements. Coupling sentiment trends with occupancy and review volume can offer actionable intelligence for both performance management and guest experience design.

Conclusion

In conclusion, this project has successfully achieved the primary objectives. The text analysis investigated trends in guest reviews and competitor descriptions, providing insights into competitor listing features and strategies, which will enable SingStay to effectively adapt to market demands. Predictive model was used to identify key factors influencing occupancy, supporting the business goal of improving occupancy rates by identifying and prioritising these influencing features. Visualizations were developed to monitor and evaluate changes in listing performance over time, ensuring consistent alignment with business objectives by tracking the effectiveness of improvements and adopting strategies as needed.

SingStay has outsourced the maintenance and support of the models used to a third party service provider, where further iterations and improvements can be made when more data is available. It can be used to assist SingStay to remain in-sync with current short-term rental market trends and latest top features that affect occupancy. With the technology-driven analysis covered in this project and the integration of visualisations into their internal system, SingStay is well-positioned for growth, with potential to improve occupancy rate, generate profit, and demonstrate the venture's value to the company within the given 5-year timeline.

References

Almirgouvea. (n.d.). The-Crisp-DM-Methodology/images/Crispdm.png at main · almirgouvea/The-Crisp-DM-Methodology. GitHub. https://github.com/almirgouvea/The-Crisp-DM-Methodology/blob/main/images/Crispdm.png

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys (CSUR), 46(4), 1–37. https://doi.org/10.1145/2523813

Grand View Research. (n.d.). *Short term vacation rental market report*. Grand View Research. https://www.grandviewresearch.com/industry-analysis/short-term-vacation-rental-market-report

Hospitable Team. (n.d.).Airbnb Ratings: A Host's Guide. Hospitable. Retrieved January 17, 2025,from https://hospitable.com/airbnb-ratings/#:~:text=The%20Airbnb%20star%20rating%20scale,on%20what%20their%20listing%20promises

Inside Airbnb. (n.d.). *Get the data - Singapore*. Inside Airbnb [Dataset]. Retrieved January 17, 2025, from https://insideairbnb.com/get-the-data/

Inside Airbnb. (2025). *Data dictionary/metadata* [Google Spreadsheet]. Google Sheets. https://docs.google.com/spreadsheets/d/1iWCNJcSutYqpULSQHlNyGInUvHg2BoUGoNRIGa6Szc4/edit?gid=1322284596#gid=1322284596

Inside Airbnb. (2025). *Data dictionary/metadata* [Google Spreadsheet]. Google Sheets. https://docs.google.com/spreadsheets/d/1iWCNJcSutYqpULSQHlNyGInUvHg2BoUGoNRIGa6Szc4/edit?gid=1322284596#gid=1322284596

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer.

Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/

Renald Yeo. 2024. Singapore's 2023 tourism receipts hit S$27.2 billion, may reach record S$29 billion in 2024 Singapore's 2023 tourism receipts hit S$27.2 billion, may reach record S$29 billion in 2024

Webb, A. R., & Copsey, K. D. (2011). Statistical Pattern Recognition (3rd ed.). Wiley.

Tan, R. (2023). Managing Business Analytics Projects: Construction, Transition, Roll-out, and Maintenance [PowerPoint slides]. Canvas

Appendix A - Links to Project Details

Project Gantt Chart

https://docs.google.com/spreadsheets/d/1-ITzaa0embkACoyCVAy_ceINJsJHzfdk/edit?gid=1539492014#gid=1539492014

Project Burndown Chart

https://docs.google.com/spreadsheets/d/1W4iWcn-s09xMvO1joPZro9CpWA1lSTLh/edit?usp=sharing&ouid=105049538854666275051&rtpof=true&sd=true

Raw Data

https://drive.google.com/drive/folders/1Un5uPD2TrTCEoO1Ep0q_9YGBRk1wiNw5?usp=sharing

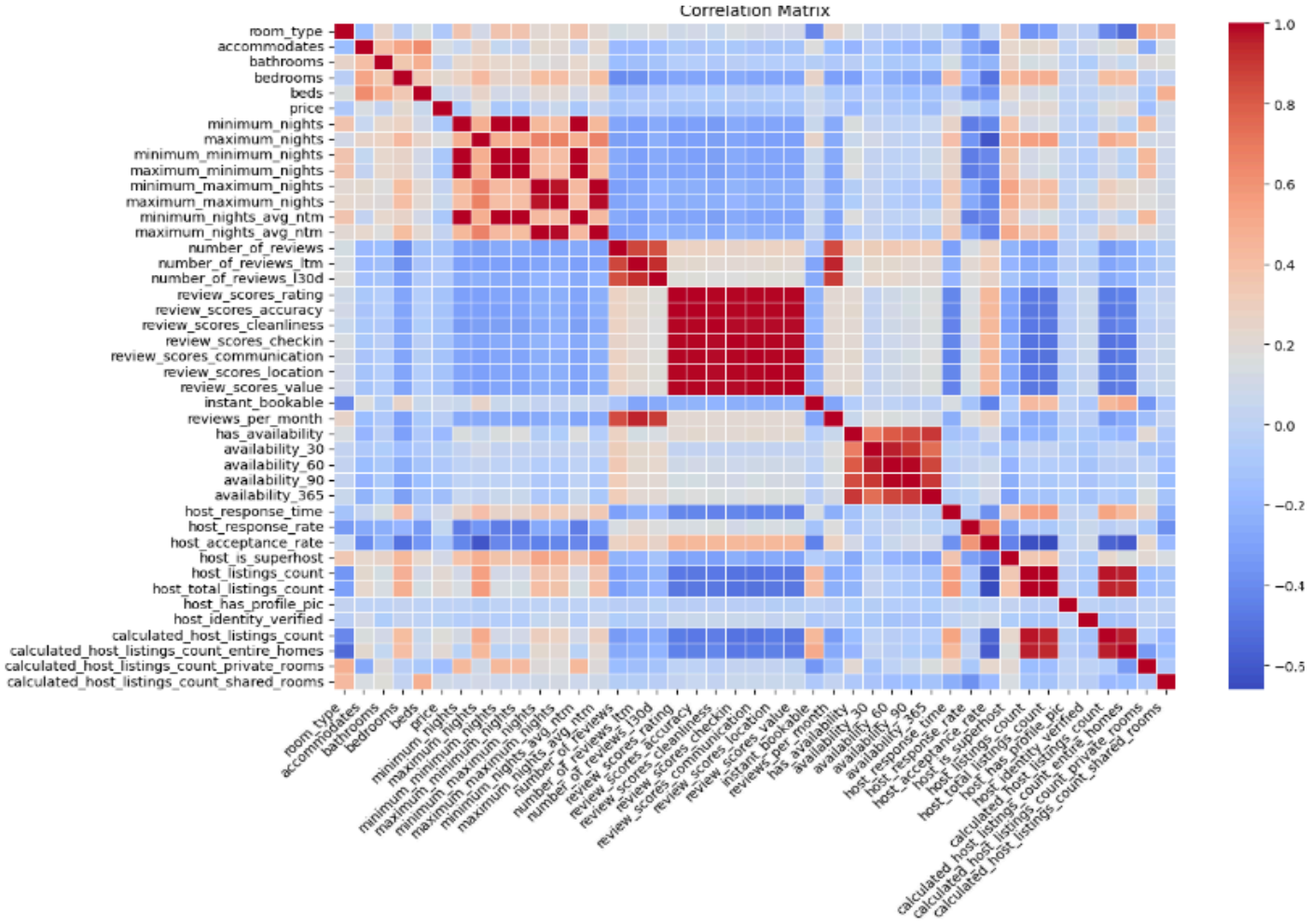
Cleaned Data

https://drive.google.com/drive/folders/1aXNYBcdE495bbETIhLq0RZbGyoyUh8O4?usp=sharing

Google Colab - Python Codes

- Data Preparation
  https://colab.research.google.com/drive/169kWdRszsDSZKan7rgO9zIRHXskjCXAl?usp=sharing
- Text Translations
  https://colab.research.google.com/drive/1NpBEkJ2hINvZIp0k0l-GMpX20mH3_TQy?usp=sharing
  https://colab.research.google.com/drive/1APggpWJMXeC0rb-hnIliDaBgllG9Pjdx?usp=sharing
- Text Exploration & Topic Modeling
  https://colab.research.google.com/drive/1aNhqqMZdHrirGZsQxkXCBoM58UDMgVCS?usp=sharing
- Sentiment Analysis
  https://colab.research.google.com/drive/1eo69d4knvawynMPG0dgysjs6LBvYdgCc?usp=sharing
- Predictive Model & Visualisations
  https://colab.research.google.com/drive/156RBE5sLhcX9QOXojh2UfGOuvPyO3ll4?usp=sharing

Appendix B - Correlation Results



Appendix C - Feature Reduction Justification

Feature Kept	Features Dropped	Reason	Method
-	'id', 'host_id', 'reviewer_id', 'reviewer_name', 'listing_id'	Unique identifiers with no analytical or predictive value	Domain Knowledge
'neighbourhood_cleansed'	'longitude', 'latitude', 'neighbourhood_group_cleansed'	Retained interpretable neighborhood-level granularity; dropped features were either too specific or too broad	Domain Knowledge
'price_band'	'price'	Raw price was skewed and affected by outliers; binned into interpretable price bands for better modeling	Domain Knowledge
'LTopic', 'RTopic'	'name_tokens', 'description_tokens', 'neighbourhood_overview_tokens', 'comments_tokens', 'host_about_tokens'	Covered by topic modeling outputs; dropped to avoid redundancy	Domain Knowledge
-	'polarity', 'subjectivity'	Redundant with topic modeling; sentiment scores not useful for model interpretability in this context	Domain Knowledge
-	'host_location', 'host_neighbourhood', 'host_verifications', 'license', 'date'	Low interpretability or limited actionable insight; difficult to use in deriving recommendations	Domain Knowledge
host_total_listings_count	'host_listings_count', 'calculated_host_listings_count', 'calculated_host_listings_count_entire_homes', 'calculated_host_listings_count_private_rooms', 'calculated_host_listings_count_shared_rooms'	Highly correlated; retained the general and interpretable aggregate	Correlation + Domain Knowledge
'maximum_nights_avg_ntm', 'minimum_nights_avg_ntm', 'maximum_nights', 'minimum_nights'	'minimum_minimum_nights', 'maximum_minimum_nights', 'minimum_maximum_nights', 'maximum_maximum_nights'	Highly correlated night-based constraints; retained averaged and standard metrics for generalizability	Correlation + Domain Knowledge
'reviews_scores_rating'	'review_scores_accuracy', 'review_scores_cleanliness', 'review_scores_checkin', 'review_scores_communication', 'review_scores_location', 'review_scores_value'	'review_scores_rating' is an aggregate of these; inclusion would introduce multicollinearity	Correlation + Domain Knowledge
'occupancy_rate' (target)	'availability_30', 'availability_60', 'availability_90','availability_365', 'has_availability'	Occupancy rate already derived from availability_365; others would leak target or create redundancy	Correlation + Domain Knowledge
'reviews_per_month'	'number_of_reviews_ltm', 'number_of_reviews_l30d', 'number_of_reviews'	Highly correlated volume metrics; reviews_per_month was retained as the most normalized and interpretable	Correlation + Domain Knowledge

Appendix D - Summary of Key Insights and Recommendations

Focus Area	Key Insight	Recommendations
Amenities	Listings with specific amenities (e.g., private entrance, workspace, TV, security cameras) have higher occupancy.	Include: Private entrance, exterior security cameras, dedicated workspace, TV to enhance comfort, privacy, and utility for work/leisure travelers.
Listings Statistics	Longer review history, higher review scores, and better host acceptance rates correlate with increased occupancy.	Build and maintain a credible review history; monitor review_scores_rating and host_acceptance_rate regularly; ensure prompt guest communication and booking acceptance.
Booking Policies	Moderate flexibility in minimum and maximum nights (especially 8–60 nights) is associated with higher occupancy. Very short or long limits reduce appeal.	Reassess and test different booking duration limits across listing types; apply A/B testing or regional configurations where needed.
Review Topics	Highly rated reviews emphasize comfort, convenience, and unique experiences (from topic modeling: RTopic_2, RTopic_9).	Highlight location and comfort in descriptions; train hosts to deliver personal touches; experiment with themed listings (e.g., Pokémon rooms, boutique units).
Location Strategy	Central Region has highest occupancy and reviews; fringe areas (e.g., Marina South, Tuas) show untapped potential despite low listing volume.	Expand into high-performing fringe areas; Maintain strong presence in the Central Region; collaborate with STB or transport providers to improve guest accessibility via shuttle services to MRT or attractions.
Review Engagement	Listings with higher review volumes, especially Superhosts, perform better. Lower-priced listings tend to attract more reviews.	Implement post-stay review requests and automated reminders; strive for Superhost status, especially in fringe regions to increase trust and visibility.
Price Positioning	Listings priced between $301–$500 showed the highest occupancy. 1-room units, though higher priced, performed well across all regions.	Use reference price bands for optimization. Maintain and promote 1-room listings for consistent performance and guest privacy value.
Performance Tracking	Performance varies by configuration. Feature changes (e.g., amenities, pricing, booking policies) should be evaluated over time.	Establish post-feature-change tracking dashboards (e.g., occupancy rate, reviews, booking lead time) to evaluate the effectiveness of updates.