Predicting AirBnB pricing for maximising profits

Project Group 3

500680807
520077942
530220947
530723345
540573758

1. Introduction

Airbnb has emerged as one of the top choices for renters due to its flexibility in pricing and date range in this highly competitive landscape. However, this popularity presents significant challenges for hosts and property investors. There are several caveats they need to think about to ensure they are maximising profits. This report aims to analyse Airbnb data from Sydney, a thriving tourist and business hub, and identify the key factors that drive rental prices. With thousands of listings in the city, hosts and property managers must not only set competitive prices but also differentiate their offerings in a crowded market. By understanding which property attributes—such as location, amenities, and property size—carry the most weight, we will build machine learning models to predict optimal daily prices.

Beyond predictive modeling, this project also aims to provide actionable insights from Airbnb data, helping hosts and investors refine their strategies in Sydney's dynamic market, where demand is influenced by seasonality, major events, and evolving travel trends.This report evaluates six machine learning models for predicting Airbnb prices: ElasticNet regression, k-nearest neighbors, random forest, gradient boosting, support vector machines, and a model stacking ensemble. Each model was chosen for specific strengths, such as ElasticNet's regularisation in high-dimensional data or gradient boosting ability to capture complex patterns. Detailed model selection rationale, evaluation, and comparison are provided in the report to highlight which model performs best for pricing optimisation in Sydney's competitive Airbnb market.

2. Problem Formulation and Objectives

In Australia's highly concentrated and competitive online travel booking industry, Airbnb is a dominant player, facing direct competition from platforms like Booking.com, Expedia Australia, Lux, and WEB Travel Group (Francis, 2024). Short-term rental markets, however, are prone to revenue volatility, pushing hosts and property managers to seek ways to maximise profits through innovative, data-driven strategies. In response, this report aims to develop a data-driven pricing strategy that identifies patterns in market demand, helping hosts and investors make informed decisions to optimise revenue. This report introduces machine learning methods used to capture complex market dynamics and provide precise pricing predictions (Soleimani, 2024).

Key questions addressed include:

- What is the optimal rental price for an Airbnb in Sydney?
- What differentiates the most successful hosts, and how do attributes like 'Superhost' status or profitability define success?
- How can hosts navigate the price-occupancy rate trade-off to maximise revenue?
- What are the primary factors influencing rental prices and occupancy rates?

Hypothesis:

1. Different host types (e.g., real estate companies vs. individual hosts) benefit from tailored pricing strategies.

The primary goals are to predict daily Airbnb rental prices, optimise pricing strategies for maximum occupancy and revenue, and offer three actionable strategies to enhance decision-making for hosts. Model performance will be evaluated using metrics like Root Mean Square Error (RMSE), and R-squared (R^2), ensuring predictions are both accurate and actionable.

3. Initial Exploratory Data Analysis (EDA)

Before conducting analysis on the features, the report identified how much missing data is present in the dataset (see Figure A1). Columns such as description, bathrooms, calendar_updated and neighbourhood_group_cleansed have nearly 100% missing values. Due to their lack of information, they have to be dropped or imputed because they do not add sufficient information. This would be further discussed in data cleaning.

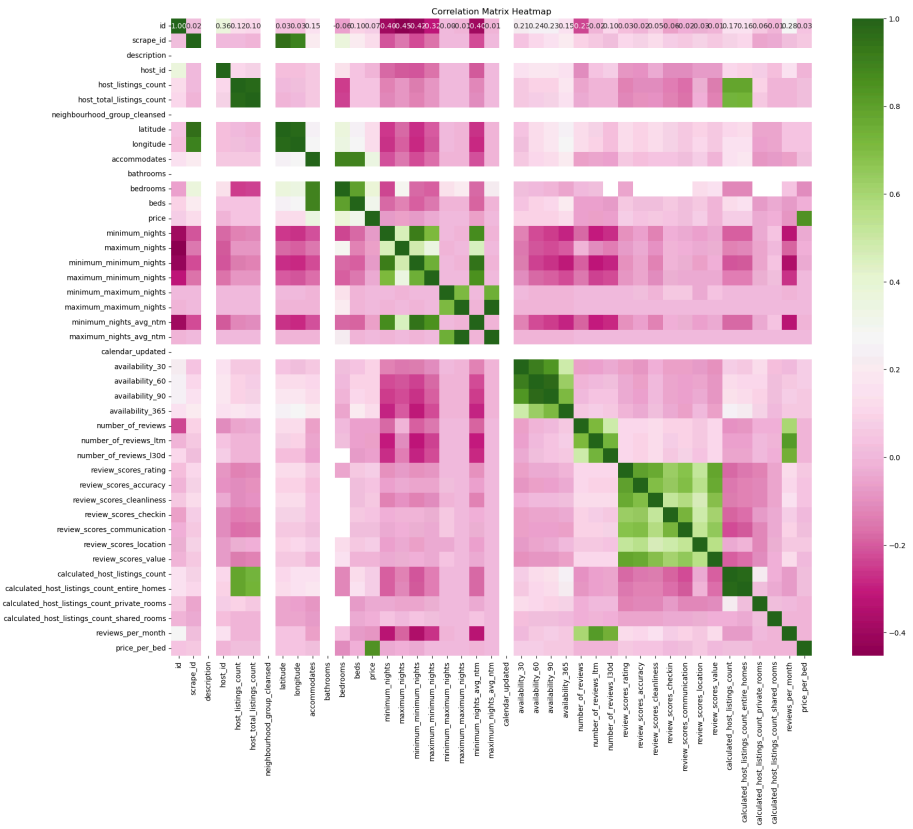


Figure 1. Correlation Matrix with Relevant Features

To understand variable relationships, a heatmap of the correlation matrix was generated, revealing key insights. High correlations were observed among bedrooms, bathrooms, and accommodates, as larger properties tend to host more guests. This suggests potential multicollinearity, warranting careful feature selection to avoid redundancy. Positive correlations among review scores (e.g., rating, accuracy, cleanliness) indicate a consistency in service quality. In contrast, weak or negative correlations with features like availability and price per bed suggest that higher availability is associated with lower rates.

Features like host_listings_count and calendar_updated also showed weak correlations, indicating limited relevance for guest satisfaction modeling.

The next steps in the EDA process would involve further investigation into potential outliers, data distribution, missing information and feature transformations to ensure that the model captures the key drivers of the target variables without unnecessary redundancy which will be further elaborated in data cleaning.

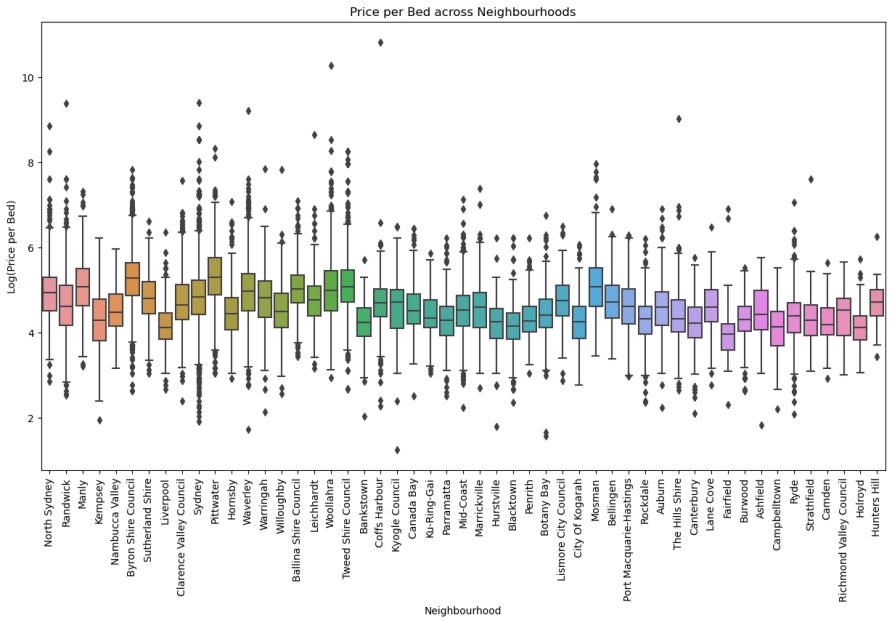


Figure 2. Boxplot to understand the Log-price distribution across neighbourhoods

Figure 2 illustrates the distribution of log-transformed price per bed across various neighbourhoods in the context of Airbnb listings, revealing several key insights. In prime locations such as North Sydney, Randwick, and Byron Shire, there is significant price variability, indicating a mix of high-end and budget listings driven by desirability and proximity to attractions. In contrast, regions like Richmond Valley Council and Hunters Hill show narrower price ranges, suggesting a more uniform property type and consistent pricing. Notably, North Sydney and Byron Shire feature higher-than-average price outliers, likely representing luxury or unique listings aimed at tourists willing to pay a premium for desirable locations. Additionally, suburbs like Ku-Ring-Gai and Kogarah exhibit lower median prices per bed compared to the more sought-after North Sydney and Randwick, which experience higher demand and, as a result, increased pricing.

The choropleth map (see Figure A2) provides the distribution of average prices among all neighbourhoods and helps identify factors that lead to the vastly varied prices.

- **Popularity:** Neighbourhoods with a high listing count, like CBD, Ashfield, and Parramatta, tend to have moderate to low prices. Maintaining a lower price can be essential as the competition increases with increasing listings. However, prices can remain high in affluent and premium neighbourhoods even with minimal listing due to an elevated demand.

- **Availability:** Neighbourhoods with greater availability but less demand tend to have low average listing prices as the properties there might struggle to fill vacancies. Premium neighbourhoods tend to have limited short-term availability, but they can attract higher prices with a higher demand.

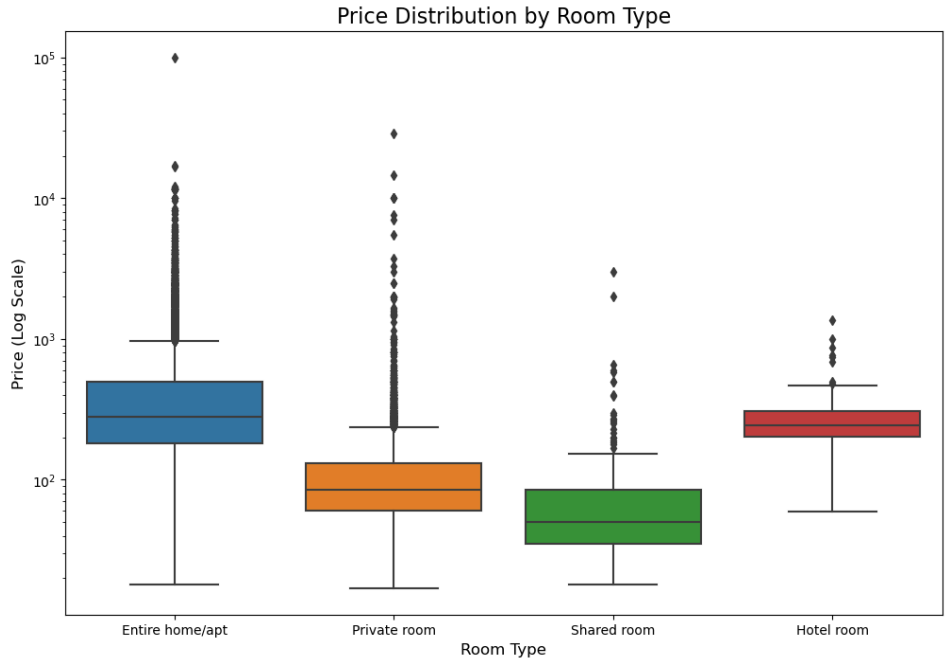


Figure 3. Price distribution by room type

The plot (Figure 3) illustrates the variation in listing prices for different property types across all neighbourhoods prior to data cleaning, enabling investors and hosts to grasp the factors that influence price variability. Entire homes and apartments showcase the widest price range and the highest average prices, indicating strong demand and larger spaces, often situated in desirable neighbourhoods. In contrast, private rooms present a lower median price, making them more affordable; however, the presence of numerous high-priced outliers suggests that factors such as neighbourhood sentiment, host behaviour, and availability can significantly affect pricing. Lastly, shared rooms represent the most affordable option, featuring the lowest median price and a very short price spread, which may indicate high availability in budget-friendly neighbourhoods managed by hosts with multiple listings.

The log-transformed price distribution across neighbourhoods reveals how availability impacts prices in high-density areas (see Figure A4). Premium neighbourhoods exhibit significantly higher average prices compared to those with lower demand, even when availability is limited. While these sought-after areas maintain consistently high demand and pricing, non-central locations do not experience substantial price increases. It is key to note that seasonal factors and neighbourhood sentiment can influence property values, with low-priced outliers in less desirable areas offering insights into pricing dynamics, particularly in premium regions with varied property availability.

The monthly distribution of Airbnb reviews is heavily right-skewed, indicating that most properties receive few reviews each month (see Figure A5). This pattern can affect pricing across neighbourhoods. Typically, guests leave negative reviews to caution others about a property, while they are less likely to write positive reviews.

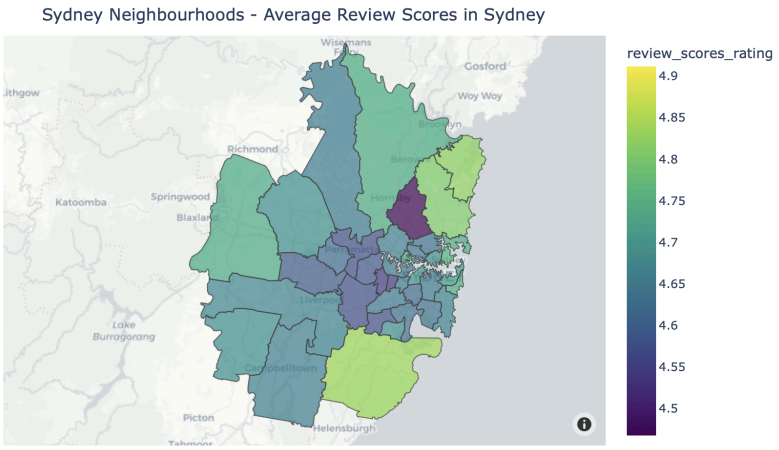


Figure 4. Average Review Score Rating across Neighbourhoods

Figure 4 indicates the Average Review ratings, which helps identify neighbourhoods with higher or lower customer satisfaction.

- **Neighborhood Affluence** : Neighbourhoods with a higher review score, like North Sydney, Mosman, and Pittwater, are well known for their scenic views, wealth, and affluent lifestyle. They frequently have immaculate homes, attracting wealthy tourists. This implies that hosts may have invested heavily in upscale lodgings, improving guest evaluations. Conversely, Inner West suburbs with budget listings and less maintenance average fewer ratings.
- **Host Behaviour**: Experienced hosts with only a few properties who provide excellent service tend to score higher. ‘Superhosts’ with a good response time are dominant factors in deciding the user review score. Contrarywise, hosts with minimal experience or multiple listings might not be able to pay the utmost attention to their properties and are found in lower-rated neighbourhoods.
- **Neighbourhood Sentiment and Property Type :** Neighbourhoods with a high review score tend to be densely populated by ‘Entire Homes,’ which suggests that these properties are used by families who tend to stay for a longer time due to the highest-rated neighbourhoods being in close proximity to beaches and restaurants.

The Top 20 neighbourhoods with the highest Airbnb listings indicate areas of dense population, with Sydney CBD as the central hub offering diverse accommodations for tourists and business travellers (see Figure A3). Other popular neighbourhoods are either centrally located or near attractions, leading to intense competition and price variability influenced by reviews. High-demand areas often experience seasonal trends, resulting in inflated prices during peak seasons. Popularity generally correlates with higher review scores, revealing pricing patterns.

4. Data Processing

4.1 Data Understanding

This dataset includes approximately 37,636 listings with 75 attributes, offering detailed property, host, and rental information. While this large dataset enhances model accuracy, it requires substantial computational resources. Key promising features identified include numerical attributes (e.g., bedrooms, bathrooms, accommodations), property type (e.g., room type), location data (latitude, longitude, neighbourhood), and host attributes (e.g., rating, verification, Superhost status). Review metrics (e.g., accessibility, review count, ratings) and sentiment analysis on "host_about" and "neighbourhood_overview" provide further insights into pricing and consumer preferences.

This dataset allows comprehensive analysis of Sydney's Airbnb market, though regional price disparities may affect generalisability, emphasising the need for data cleaning. Key features such as property type, location, and host status directly address questions on rental price drivers and enable accurate price trend predictions. Feature selection or dimensionality reduction will be applied to prevent model overfitting.

4.2 Data Cleaning

During the initial Exploratory Data Analysis (EDA), we identified several columns that were either irrelevant or contained too much missing data, limiting their contribution to the model's performance (as seen in Figure A1). As a result, these columns were dropped to streamline the dataset and ensure the model focuses only on valuable information.

The columns removed include:

- **Metadata-related columns**: These columns, such as "calendar_updated", "last_scraped", "scrape_id", "listing_url", "calendar_last_scraped", and "source", provided logistical or scraping details that do not impact rental price predictions.
- **Media and URL links**: Columns like "host_picture_url", "host_thumbnail_url", "host_url", "picture_url", and "host_verifications" were removed because media links and URLs are not informative for predicting rental prices.
- **Text-heavy and redundant columns**: Fields such as "description", "host_about", and "neighborhood_overview" were dropped as their sentiment values were already incorporated into the model through sentiment analysis which will be further explained in the feature engineering section. Including the raw text would have been redundant.
- **Geographical and property-related columns**: "neighbourhood_group_cleansed", "neighbourhood", "name", "bathrooms_text", "bedrooms", "bathrooms", "host_since", "first_review", and "last_review" were also removed either due to redundancy or high levels of missing data that made them less useful for the model.
- **Has availability columns:** "'has_availability', 'availability_30', 'availability_60', 'availability_90', 'availability_365' were dropped to avoid data leakage. These fields reveal future booking data, giving the model an unfair advantage by allowing it to "see" future demand, which wouldn't be known at the time of setting the price. Dropping these columns ensures the model relies on legitimate factors like property features and location to make unbiased, reliable predictions.

Additionally, several data cleaning steps were implemented to ensure the dataset is of quality for analysis:

1. **Normalising text data:** All the text data was converted to lowercase to avoid bias during the process of sentiment analysis. Leading and trailing spaces were also removed to ensure consistency and prevent errors caused by inconsistent formatting.
2. **Standardising values:** In columns like “license”, missing entries were filled with “None” for consistency. License numbers were simplified to retain only essential information (e.g., "PID-STRA"), converting them into a more usable categorical variable.
3. **Boolean values:** Boolean values of “yes”/”no” were changed to “True”/’False” values making them easier for machine learning algorithms to interpret and process effectively.
4. **Duplicates:** Duplicates were identified and removed to avoid bias in the model’s predictions. Duplicate data could skew the model’s results or even introduce data leakage, inflating model’s performance artificially.
5. **Deleting incomplete rows:** Rows with missing price (target variable) were dropped as they could not contribute to the training and testing the model, given that the price prediction is the core goal of this analysis.

5. Feature Engineering

By analysing the data from EDA, we have introduced some feature engineering to improve the accuracy of model’s predictions.

1. **Sentiment analysis:** We applied sentiment analysis to text-based inputs such as “host_about” and “neighbourhood_overview”. This allowed us to convert the descriptions into numerical sentiment (positive or negative). By doing so, we could assess if the tone of these descriptions had any influence on AirBnB pricing. For consistency purposes, missing or unknown entries were assigned a neutral score of 0. This approach enables the model to account for qualitative aspects of a listing, which may impact customer perception and ultimately, the rental price.
2. **Splitting complex data:** It was identified that some columns had multiple pieces of information, thus, we split them into more focused parts to improve clarity.
    a. "bathrooms_text" was divided into "bathroom_number" (the quantity) and "bathroom_type" (e.g., private or shared), making it easier for the model to interpret the details.
    b. The "host_verified" column was broken down into specific categories like "host_phone_verified", "host_email_verified", and "host_work_email_verified".
    c. Data columns were split into year, month and day to capture better trends in relation to seasonality.
3. **Log transformation of pricing:** To address the high variability in rental prices, we applied a log transformation to the response variable. This stabilises variance and normalises the price distribution, enhancing the model’s performance by reducing the impact of outliers and enabling more effective learning of pricing patterns.

6. Post-data cleaning EDA

The approach taken in this EDA section aims not to directly see which variable directly influences price but rather continue on finding patterns in the dataset and identifying underlying relations.

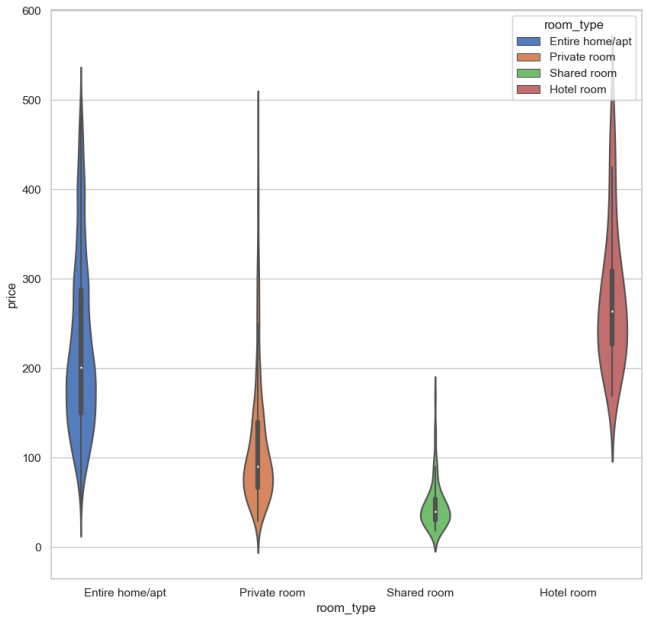


Figure 5. Distribution of Price for Different Room Types

The plot illustrates the variation in listing prices for different property types across all neighbourhoods after data cleaning leading to insightful changes in the chart.

- **Reduction in Outliers:** it can be observed in the violin chart that extreme values of the outliers are much less pronounced as data cleaning required dealing with these outliers and removing any listings with missing price. It also illustrates a varied distribution.
- **Accuracy in Price**: The violin chart has a more normally (evenly) distributed chart as data cleaning helped in removing any skewness or biases in the data which produced visualisations not reflective of the market behaviour.
- **Private Rooms:** although few outliers still remain, there is a large concentration In the lower half of the price points indicating affordable options with a few exceptions based on neighbourhood and ratings.
- **Hotel Room :** a significant change can be observed in the distribution of the violin chart as the data for hotel rooms is much more distributed with the highest median price. This signifies the location of the majority of hotels being in neighbourhoods with high popularity and hotels being near tourist attractions and cities where popularity is excessive. Hotels that are professionally managed also add a layer of premium feel and satisfaction, justifying the higher price.

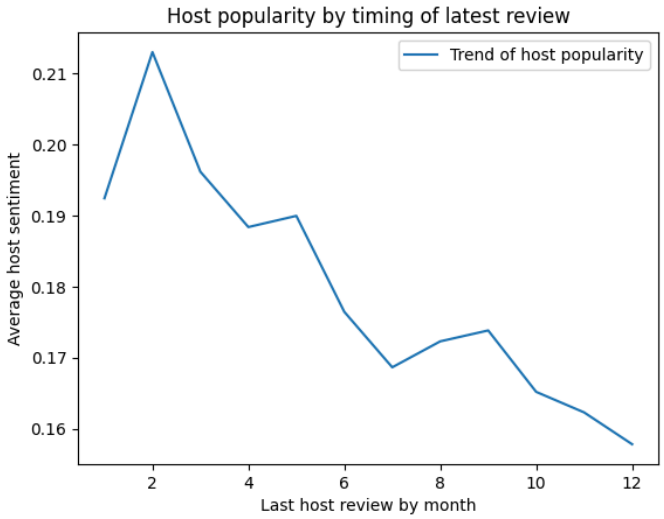


Figure 6: Popularity over cleaned last review day grouped into months

This plot shows a decreasing trend after 2 months of no review updates. This indicates that to remain a well-liked and subsequently competitive host in AirBnB, it is important to have regularly updated reviews with a period of no longer than 2 months.

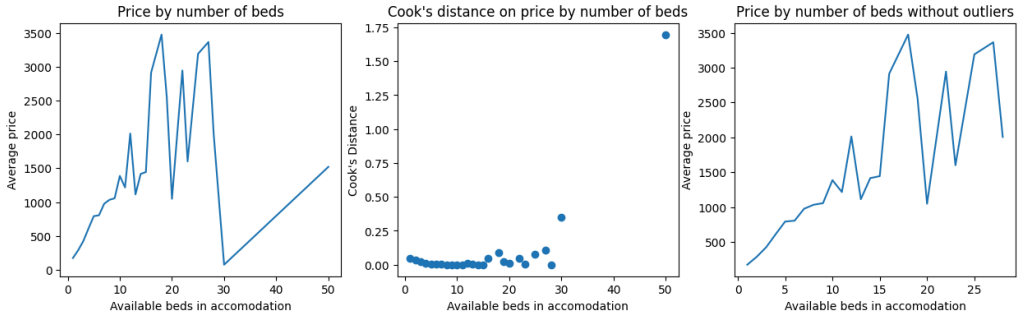


Figure 7: Average price of listings over number of beds

From the initial EDA we only see reasonable correlation with revenue and accommodation against price. With price being the output this assignment is focusing on, it is worth investigating price against some features of accommodations, one of which is the number of beds. The plot on the left and in the middle represents the average price with all data points. As the price on accommodations with 30 and 50 beds fluctuate immensely, The relative distance (Cook's distance) was measured iteratively and both data points were found to have said distance higher than 1. After removal of such data points, the plot on the right shows an updated correlation, showing significantly lesser variation (with an R-squared value of 0.647).

Initial EDA indicated price variation in certain outlier neighborhoods and prime locations. Further analysis, using latitude and longitude on log-scaled prices, revealed a V-shaped scatter plot (see Appendix A6). Given the limitations of traditional R-squared, cosine similarity was used, showing a mean similarity

>0.95. This suggests a reflective price trend: accommodations are more costly near the city center or in distant, scenic areas like Byron Shire.

7. Methodology (Model Building)

In this report, a total of six models have been used; elastic-net linear model, k-nearest neighbour, random forest, gradient boosting, support vector machines and model stacking. For distance sensitive algorithms (k-NN and SVM), standardisation to 0 mean and unit variance and Principal Component Analysis (PCA) is used covering 95% cumulative variance, reducing 59 variables to 34.

7.1 ElasticNet Linear Model

The linear model chosen is ElasticNet regularised-linear regression. Compared to normal linear regression, ElasticNet uses a combination of Lasso and Ridge regularisation to penalise high-complexity variables. The advantage of using a linear model lies in its interpretability and inference time: each variable coefficient can be optimised and extracted to form an equation for the expected value of an entry, with error of mean 0 and finite variance. Once the model is trained, only these coefficients are kept and the prediction can then be calculated without the use of the training data.

Lasso regularisation uses the L1 norm and Ridge regularisation uses the L2 norm in their penalty. Using a regularisation method in risk minimalisation prevents the potential issue of overfitting. There is no general ‘decision’ as to which form of regularisation between Lasso and Ridge yields better results, particularly when it comes to linear regression as restricting covariance will always result in the loss of R-squared. Therefore, only a small penalty and a small ratio of L2 is used. Any hyperparameter tuning will recommend using an extremely small penalty coefficient and ElasticNet ratio.

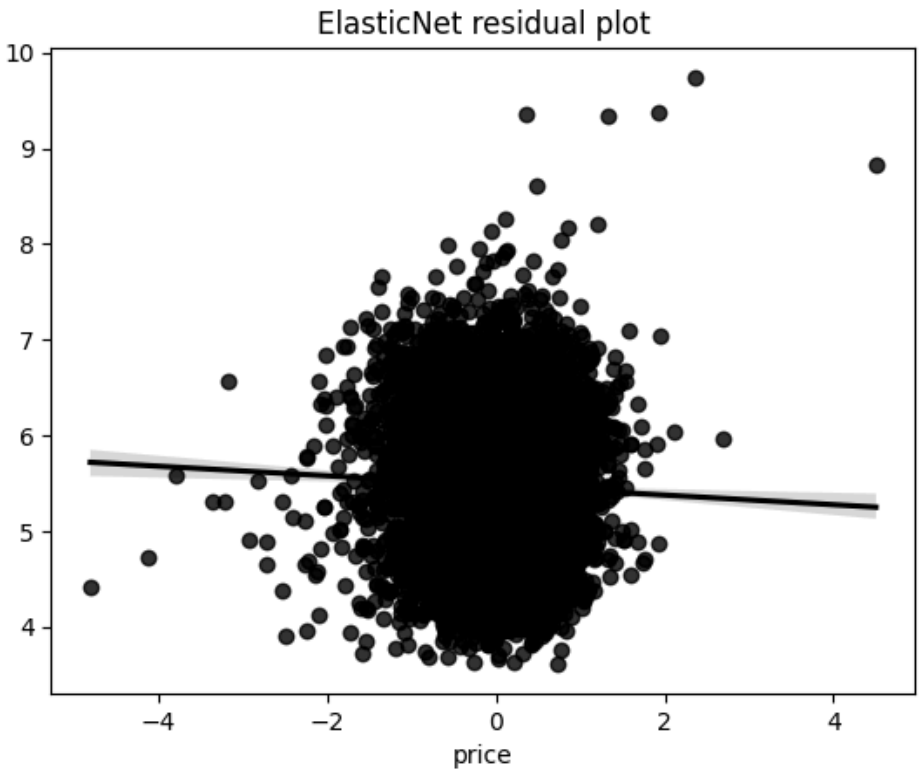


Figure 8: Residual plot of ElasticNet regression.

Figure 8 represents the residuals from the ElasticNet Linear model. Residuals seem to be clustered around the center depicting a non-constant variance possibly indicating heteroskedasticity. This graph also shows high outliers indicating potential large prediction errors.

A validation set was used here using train-test-split. The ElasticNet Linear model returns an RMSE of 0.544 and an R-squared of 0.629. However because of potential heteroskedasticity, other models will also be further explored and analysed for predictions.

7.2 k-Nearest Neighbour

The k-Nearest Neighbour (hereon referred to as k-NN) algorithm is a regression model using similarity between observations. Each prediction is made using data within the training dataset: the observations with the highest similarity are taken and averaged arithmetically to obtain the regression result. Similarity is calculated using a metric with each prediction entry as the center of the cluster. The number of neighbours is controlled as a hyperparameter. The advantage of k-NN is its lack of training and its flexibility: there is no function to be trained, and does not assume any function form on its training data either. This makes k-NN incredibly adaptable to nonlinear data. The lack of training is also known as a ‘lazy learning algorithm’, making the method suitable for noisy data as they do not affect the result of the output as much as other models would.

Whilst it is theoretically possible for hyperparameter tuning on k-NN, it is infeasible. Using any other metrics (such as Canberra metric) outside of the given options on distance and weight calculation will require a separate, external wrapper function to be called each and every time a test data entry is regressed. As a result, the regression model slows down dramatically even in parallel computing, as the additional wrappers often compute very slowly. After cross-validated optimisation, the **k** coefficient that gives the best performance was set at 13. The optimisation of the **k** coefficient is essential. Using a small **k** allows for very little choice of neighbours and will likely overfit with great bias; while using a large **k** will include too many samples and increase variance in the regression.

The non cross-validated model returns an RMSE of 0.500 and an R-squared of 0.686. The 10-fold cross-validated model returns an RMSE of 0.499 and an R-squared of 0.689.

7.3 Random Forest

The random forest model is used because of its strong predictive capability on complex, non-linear datasets. As a non-parametric, non-linear ensemble method, it combines predictions from multiple decision trees using bootstrapping (sampling with replacement) and bagging (aggregating results across diverse tree structures). These techniques effectively reduce model variance, making Random Forest highly resistant to overfitting and capable of generalising well to test data (Breiman, 2001). To further enhance the model's performance and reliability, hyperparameters were carefully tuned in this report with specific goals of optimising both accuracy and computational efficiency, as follows:

Hyperparameter tuning:

1. **Parameter Space Design**
    a. **Integer Parameters:** For integer-based parameters, such as max_depth and n_estimators, we specified a random integer range. With this limitation, we prevent the need for exhaustive searching, keeping the computational demands low without sacrificing the effectiveness of parameter exploration.
2. **Optimisation Approach - Randomised Search with Cross Validation**
    a. **Randomised Search:** Given the high dimensionality of the parameter grid, the randomised search was chosen over grid search. This samples a subset of parameter configurations, which reduces computational load while exploring a broad range of values. This made it efficient and reliable for optimising larger parameter spaces (Bergstra & Bengio, 2012).
    b. **Cross-Validation:** By splitting the dataset into k-folds (in this case, 5 folds) and evaluating each parameter configuration on these subsets, the cross-validation ensured the model's robustness, reducing overfitting (Varma & Simon, 2006).
3. **Parameter Grid with Bootstrapping Option**
    a. **Parameter Ranges:** Each parameter was assigned a range with minimum and maximum values to focus tuning on effective configurations and enhance computational efficiency. By defining these ranges, we minimise unproductive explorations of the parameter space, leading to a more efficient tuning process (Probst et al., 2019).
    b. **Bootstrap Sampling Option:** In each iteration, the model was given the choice between using bootstrapped or non-bootstrapped samples. Bootstrapping is typically favoured for

enhancing model robustness and reducing overfitting, but non-bootstrapped sampling was also explored to identify the optimal sampling strategy for this dataset (Breiman, 2001).

Results and cross-validation:
Cross-validation was employed to enhance the training dataset, resulting in improved model evaluation metrics. Through this approach, the model led to cross-validated RMSE of 0.433 and an R-squared of 0.767. Since the cross-validation was run on the training dataset, lesser data points were available to the model for training. To check for model capability on the entire dataset, a validation set approach was employed which led to an RMSE and R-squared are of 0.419 and 0.780, respectively. The validated set RMSE is slightly lower, which is most likely due to the presence of extra data points available to the model to train on.

Feature importance:

In this model, Airbnb rental prices are most influenced by the "accommodates" feature, which reflects a property's guest capacity and holds an importance score of 0.30 (see Figure B1). This high score indicates that properties accommodating more guests command higher prices, particularly appealing to families and groups. Secondary features like bathroom count further support value by enhancing comfort, suggesting that hosts can strategically focus on guest capacity and amenities to boost prices and stand out in Sydney's competitive Airbnb market.

7.4 Gradient Boosting

XGBoost is a tree-based method rooted in Gradient Boosting that uses a differentiable loss function to iteratively minimise error from weak learners, ultimately creating a robust, generalised model. By training thousands of models on various subsets or trees of the training data, XGBoost can identify the optimal model that performs well not only on the training set but also on unseen test data.

Hyperparameter tuning:
To find a balance between capturing data patterns and avoiding overfitting, hyperparameter tuning is essential, especially for tree-based models like XGBoost. It helps controlling the depth of trees and improving model performance. The following considerations were made:

- Due to the extensive number of hyperparameters being tuned, the size of the grid to be searched is vast. For continuous values, the parameter space is constructed as a Normal distribution. For discrete values such as the maximum depth of the trees, a random integer value between a predefined set of values is used.
- RandomSearch with Cross Validation was chosen over Grid Search as the latter is not viable due to the size of the grid being searched and the computational requirements for it. Cross Validation was used for 2 main reasons. Firstly, to ensure that the model does not overfit on the dataset. Secondly, to obtain an overall performance of the model on the entire dataset as all folds will be tested independently and the final model metric will be derived from all the training set data.

The final features list is chosen from the best model produced with hyperparameter tuning. The best model led to an non cross-validated RMSE of 0.417 and an R-squared of 0.786.

Results and cross-validation:
Cross-validation was employed here for the same reason as the Random Forest model. XGBoost has a cross-validated RMSE of 0.417 and a cross-validated R-squared of 0.782 which are very close to the non cross-validated ones. This suggests that the model generalises well to unseen data as well as training data.

Feature importance:
The XGBoost feature importance analysis (see Figure B2) highlights room_type as the most influential factor in Airbnb pricing, with an importance score of 0.4. This underscores the strong impact of property type on price, especially for entire homes and apartments, which command a higher price due to their desirability and spaciousness. Hosts aiming to maximise income can capitalise on this insight by offering diverse room types or refining their pricing strategies to reflect the premium potential of certain accommodations—larger spaces, for instance, generally justify higher rates. Shared rooms, by contrast, emerge as budget-friendly options with limited scope for premium pricing adjustments.

7.5 Support Vector Machines

Support Vector Machines and Support Vector Regression (hereon referred to as SVM or SVR) is a relatively novel algorithm. Compared to traditional binary SVM classification where the hyperplane of maximum margins is created as separation between labels, in SVR an extra term of ‘epsilon’ is introduced with the objective that the chosen kernel function predicts a value Y that deviates no more than plus or minus epsilon on the training dataset. At the same time, a convex optimisation problem is solved to ensure the solution chosen maximises the flatness of the kernel function, akin to maximum margin. (MathWorks, n.d.)

SVMs are very resilient to outliers due to the nature of maximum margin, choosing only select data points for their support vector, therefore an outlier is unlikely to be chosen. As a non-parametric model, not only does SVM not assume linearity but is often able to find relationships between variables to support vector building.

Hyperparameter tuning on SVR on its current implementation is highly computationally expensive. SVM cannot be trained in parallel, suggesting its inability to share workload to other available computing units. This is moreso reflected in SVR. Whilst choosing a different epsilon does not have a great effect on the model and its tuning process, changing the C constant in SVR increases the training time and memory allocation to the point where the training process fails because of lack of memory. A linear change in C is an exponential change to the amount of data points iterated for the epsilon limitation. As such, the SVR model is not tuned.

In this model, a validation set was used using train-test-split. The model returns an RMSE of 0.463 and an R-squared of 0.731.

7.6 Model Stacking

Model stacking involves combining multiple individual models to improve prediction accuracy. Through these predictions, a better performing model can be created. However, the computation cost of model stacking is much higher than every individual model and also poses the risk of overfitting.
The models chosen for evaluation are from above, including k-NN, XGBoost and SVR. Their tuned parameter is copied over to ensure their best performance. The final decision model uses a Bayesian Ridge linear model, where instead of an L2 norm a high-rank pseudoinverse on the theory of generalised matrix inverses is used for regularisation. From the three models chosen from evaluation, the prediction that is the closest to the output of the decision model will be chosen as the final result.

In this model, a validation set was used using train-test-split. The model returns an RMSE of 0.458 and an R-squared of 0.737.

8. Model Evaluation and Results

To assess model performance and identify the best predictive model, various evaluation metrics were employed.

Model	**RMSE**	**R-squared (R^2)**	**RMSE Comparison (%)**	**R^2 Comparison (%)**
XGBoost	0.417	0.786	-	-
Linear ElasticNet	0.544	0.629	30.46	-19.97
k-NN	0.500	0.686	19.90	-0.38
Random Forest	0.419	0.780	0.48	-0.76
SVR	0.463	0.731	11.03	-7.00
Model Stacking	0.458	0.737	9.83	-6.23

Table 1. Model Comparison

Based on Table 1, the XGBoost model is the preferred choice for predicting Airbnb rental prices in Sydney, achieving the lowest RMSE (0.417) and the highest R^2 (0.786). The low RMSE indicates that XGBoost minimises average prediction error, with log-price predictions deviating by approximately ±0.417, suggesting closer alignment with actual prices compared to other models. Additionally, XGBoost's higher R^2 captures 79% of rental price variance, showcasing its ability to generalise well across the dataset. This combination of low prediction error and high explanatory power underscores XGBoost's effectiveness in modelling Sydney's complex rental market. In contrast, models like ElasticNet and k-NN, with higher RMSE and lower R^2, are less capable of capturing the intricate feature

interactions present, solidifying XGBoost as the optimal model choice.

The XGBoost model, identified as the optimal choice for Airbnb pricing, reveals that room type is the most influential factor, with "Entire home" or "Apartment" options commanding premium rates over shared or private rooms. Hosts can leverage this insight by pricing entire properties higher for exclusivity, while pricing shared spaces more competitively. Guest capacity, another key factor, shows that larger accommodations generally attract higher rates, making it beneficial for hosts with spacious properties to target groups or families, particularly during peak seasons. Amenities like additional bathrooms also justify higher pricing, appealing to guests seeking comfort. For properties near popular locations in Sydney, the model suggests higher seasonal prices to capitalise on demand. By dynamically adjusting pricing based on competitor data and market trends, hosts can set optimal rates that maximise occupancy and revenue, using XGBoost's accurate predictions and feature insights to stay competitive in the rental market. Given XGBoost's strong generalisation capabilities, the data quantity requirements are less critical compared to models like Random Forest, making XGBoost a superior choice for better performance.

As previously stated in Section 7.4, accommodation type is a major factor in Airbnb pricing, with "Entire home" and "Apartment" listings commanding higher rates than shared rooms. Guest capacity and amenities like additional bathrooms also increase rates, attracting groups and families. For properties near Sydney attractions, the model suggests seasonal price adjustments to maximise demand. By aligning rates with market trends and competitor data, hosts can leverage XGBoost's predictions to boost occupancy and revenue.

Across the various models, feature importance was analysed to identify which attributes most significantly impacted model performance. While XGBoost identified room_type as the top feature, this is not definitive; a more suitable model might explain the remaining 20% of variance more effectively, potentially revealing additional insights.

Limitations

While the XGBoost model explains 78.6% of the variability in rental prices, 21.4% remains unexplained, indicating potential for further refinement. Exploring alternative models could yield better predictive accuracy, although any results should be approached with caution due to possible model limitations that may still persist.

9. Data Mining: What are the best hosts doing?

Based on our regression modelling, it was identified that room_type significantly influences the price of Airbnb listings; however, this does not necessarily imply that hosts with higher-priced listings are automatically classified as "Best Hosts." To better understand what distinguishes high-performing hosts, our clients—hosts and real estate investors—are particularly interested in traits that correlate with high prices, guest satisfaction, and optimal financial returns. Thus, we define "Best Hosts" as those who exhibit attributes associated with elevated prices and performance metrics, leveraging various data-mining techniques to uncover patterns and actionable insights. Through host categorisation, a classification model

combined with feature importance analysis, Generalised Sequential Pattern (GSP) mining, and quantitative insight extraction, we can illustrate the key strategies employed by these successful hosts.

Defining “Best Hosts”

Several attribute criteria were used to categorise hosts as "Best Hosts," which helped differentiate them from regular hosts. Exploratory Data Analysis, coupled with an extensive understanding of the data, helped group the most important features.

- **SuperHost Status:** A sign of a good host, as they have more visibility and earning potential with an average review score of more than 4.8, high customer engagement and fast response times. Only a small percentage of the entire Host population makes up the SuperHost category (see Table C5).
- **Average Review Score:** A minimum review score of 4.7 to ensure high guest satisfaction
- **Host Response Rate:** Hosts who respond to 90% or more of their customers in a quick time (Figure C3).
- **Price:** Hosts with their listings in the top 20% of pricing and revenue are also considered 'Good Hosts' as they allow for income maximisation and optimisation.

Classification Model Analysis

To address the business need, this report develops a classification model using a Random Forest classifier to distinguish “best” hosts from “average” ones based on key attributes. This model effectively identifies characteristics associated with high-performing hosts and predicts whether a host qualifies as a “Good Host” with high accuracy. The Random Forest Classifier was selected for its strong performance with complex data, ability to prevent overfitting, and interpretability via feature importance metrics. This approach allows for deriving quantitative insights into the factors contributing to a host's success on the platform.

This model demonstrates a strong accuracy score, confidently classifying most hosts as “best” or “average” (see Table C1). Its high precision reduces false positives, while robust recall ensures that key characteristics of good hosts are not overlooked. Additionally, the high F1 score indicates effective detection of leading hosts with minimal misclassification. Together, these metrics confirm the model’s effectiveness in identifying and predicting important features and hosting strategies.

Feature Importance

The analysis of the feature importance derived from the Random Forest Classifier puts forth the Superhost status as the most important variable (see Figure C2) showing how strongly it associates with host performance due to increased visibility and guest trust. The average review score and listing availability follow closely which are indicative of the requirement for flexible booking options and overall guest satisfaction. Other factors like price, response rate, location (neighbourhood) and the flexibility in guest stay through minimum nights also provide a clear framework for the prioritisation of improvements needed by hosts.

Generalised Sequence Pattern (GSP)

The Generalised Sequence Pattern (GSP) analysis revealed that top-performing Airbnb hosts share key attributes: high review scores, flexibility, Superhost status, premium pricing, and fast response times. These factors enable them to achieve medium to high price points and optimal revenue (see Table C2). In

premium suburbs like Sydney CBD, Mosman, and Pittwater, successful hosts capitalise on high demand and longer-stay options to attract families and tourists, ensuring high occupancy and returns (see Figure C4). Comparative analysis showed that these hosts achieve greater revenue and review scores through prompt communication, strategic pricing, and exceptional guest satisfaction, in contrast to average or newer hosts.

This report examines the key characteristics, behaviours, and strategies that set top hosts apart from average ones and identified the following:

1. **Occupancy and Review Scores**
   Superhosts have a slightly lower occupancy rate of 58% compared to 66% for non-Superhosts, yet they achieve a higher average review score of 4.51 versus 3.36 (Table C5). This suggests that while Superhosts may attract fewer guests, they deliver superior guest satisfaction, which is vital for repeat business and a strong reputation.

   Consequently, hosts and real estate investors should prioritise customer retention over acquisition, as retaining existing customers is generally more cost-effective and can lead to increased revenue. By focusing on enhancing guest experiences, Superhosts can encourage repeat bookings and strengthen their market position.

2. **Dynamic Pricing Strategies and Their Impact on Revenue**
   Hosts in high-demand premium neighbourhoods implement dynamic pricing strategies during peak seasons to optimise occupancy and revenue while ensuring high guest satisfaction. Observations reveal that hosts in the top 20% of pricing grades achieve approximately 30% higher revenue than those in lower pricing tiers, with average prices in these neighbourhoods reaching AUD 605 compared to AUD 320 in more affordable areas (Figures C4 and A2). This indicates that elevated pricing is sustainable in high-demand sub-markets, provided hosts meet guest expectations regarding quality, location, and amenities (Figure C1).

   Therefore, for hosts and real estate investors, the emphasis should be on strategic pricing aligned with property quality and location. By leveraging insights on market demand and competitive pricing, they can maximise revenue potential, enhance occupancy rates, and solidify their position in the rental market.

3. **Superhost Status and Host Reliability**
   Airbnb Superhosts prioritise guest experiences and respond to guest inquiries promptly to improve their conversion rates and guest satisfaction. Data indicates that Superhosts earn around 15% more due to higher occupancy rates, as their listings enjoy enhanced visibility and increased trust among potential guests. With an impressive average review score exceeding 4.8 and response rates typically above 90%, Superhosts effectively foster positive guest interactions that lead to more bookings (Table C5). Moreover, Superhost listings average 182 days of availability, catering to guests' desire for flexibility in their accommodation options.

Hence, hosts and real estate investors should actively pursue Superhost status by enhancing their guest engagement strategies and ensuring swift communication. By consistently delivering exceptional guest experiences and maintaining high availability, they can significantly increase occupancy rates and revenue. This approach not only solidifies their competitive position in the Airbnb market but also builds a loyal customer base that can lead to long-term success.

4. **Flexible Availability**
   Flexible stay options, such as lower minimum nights and longer maximum stays, give hosts the opportunity to cater to a wide range of guests, from business travellers to families on vacation. By emphasising quality experiences, these hosts consistently achieve high review scores, focusing on factors such as cleanliness, value for money, ease of check-in, and prime locations. Listings experience a notable 25% increase in bookings once their average review score reaches 4.8 or higher, showing minimal variability in this trend. Exceptional reviews enhance the appeal of these listings, but offering flexible availability—especially during peak tourist seasons—truly broadens their market reach (Table C2).

   As such, hosts and real estate investors should consider adopting flexible booking strategies to maximise occupancy. By providing options for both short-term and extended stays while maintaining high service standards, they can attract a wider audience. This approach not only boosts occupancy rates—evidenced by the 85% average for highly rated hosts compared to just 60% for lower-rated ones (Table C3)—but also positions them advantageously in the competitive rental market.

Based on this comprehensive analysis of Airbnb data in Sydney, several key conclusions can be made to guide hosts and property investors. Firstly, while room type significantly influences pricing - particularly with entire homes and apartments commanding higher rates - superhost status is critical for maximising revenue as it enhances visibility and guest trust. Implementing dynamic pricing strategies during peak seasons allows hosts in high-demand areas to increase revenue by up to 30% compared to lower-priced competitors. Additionally, flexible availability attracts a broader range of guests, boosting bookings. As such hosts can leverage these insights by prioritising the maintenance of their Superhost status through excellent guest service and prompt communication to enhance their listing’s attractiveness. By implementing the recommended strategies, hosts can cater to a wide range of traveller preferences, enhance their financial performance, and simultaneously provide exceptional experiences that cultivate guest loyalty.

References

Anonymous (n.d.). Understanding Support Vector Machine Regression. *MathWorks R2024b*. https://au.mathworks.com/help/stats/understanding-support-vector-machine-regression.html

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization Yoshua Bengio. *Journal of Machine Learning Research*, *13*, 281–305. https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf

Breiman, L. (2001). *Random Forests*. https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

Fahey, J. (2023, November). *Online Travel Bookings in Australia*. Services.ibisworld.com. https://my.ibisworld.com/au/en/industry-specialized/od4163/at-a-glance

Probst, P., Wright, M. N., & Boulesteix, A. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(3). https://doi.org/10.1002/widm.1301

Soleimani, F. (2024). Dynamic Competitor Analysis and Pricing Strategy Development Using Machine Learning Models. *International Journal of Industrial Engineering and Construction Management (IJIECM)*, *2*(1), 1-10. https://www.ijiecm.com/index.php/ijiecm/article/download/12/9

Varma, S., & Simon, R. (2006). Bias in Error Estimation When Using Cross-Validation for Model Selection. *BMC Bioinformatics*, *7*(1), 91. https://doi.org/10.1186/1471-2105-7-91

Appendices

Appendix A: Exploratory Data Analysis

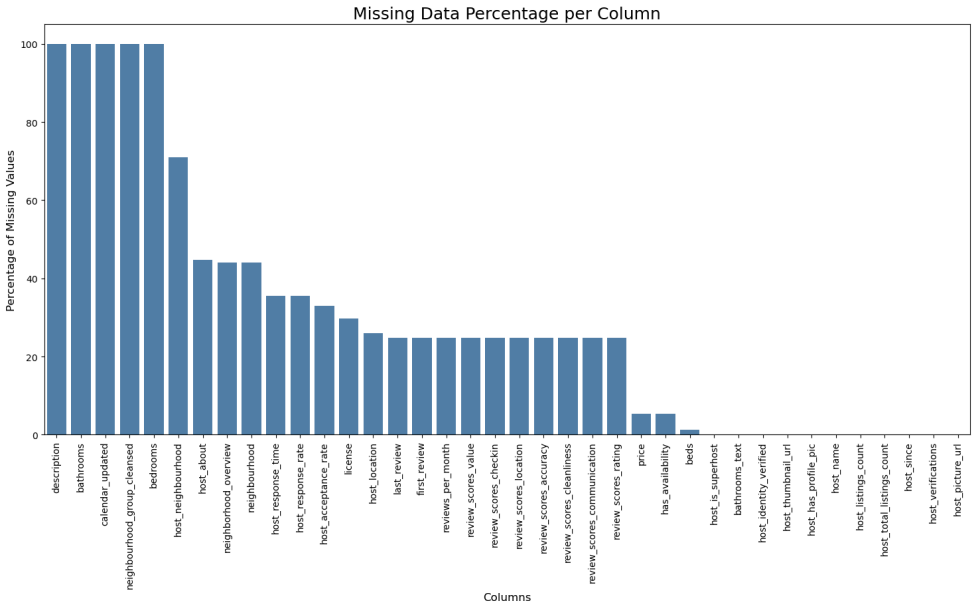


Figure A1. Identifying missing data percentage %

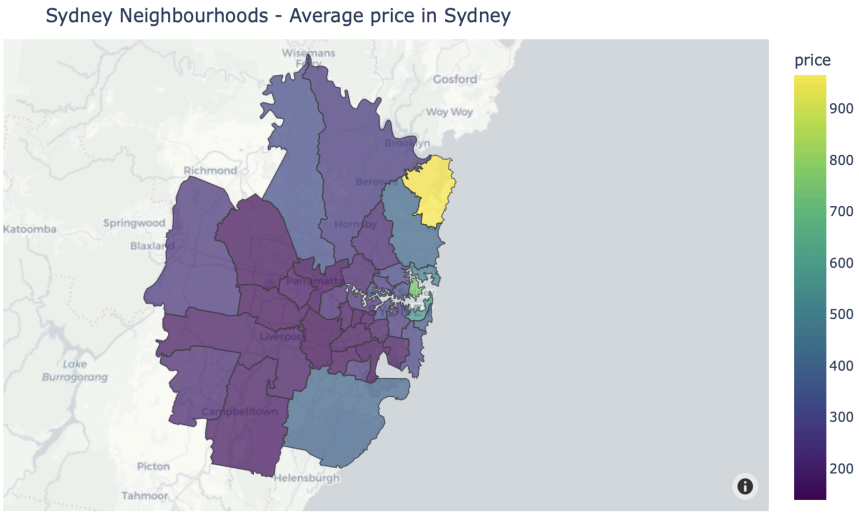


Figure A2. Average Listing Price across Neighbourhoods

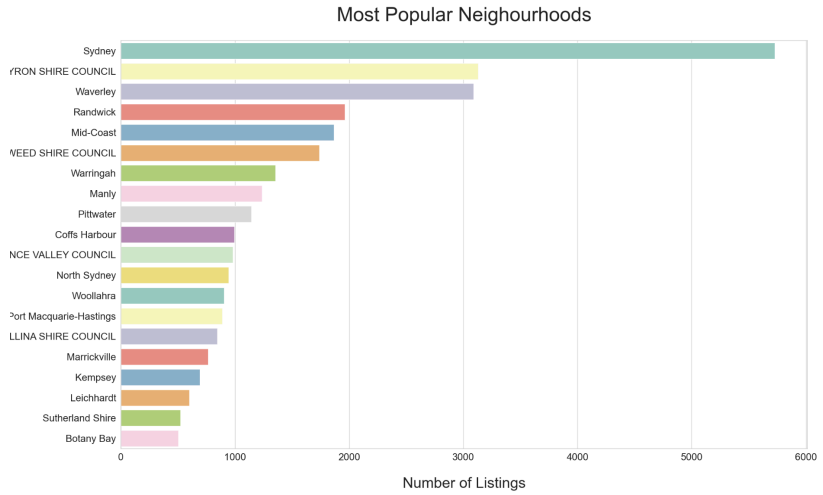


Figure A3. Highest Number of Listings among Neighbourhoods

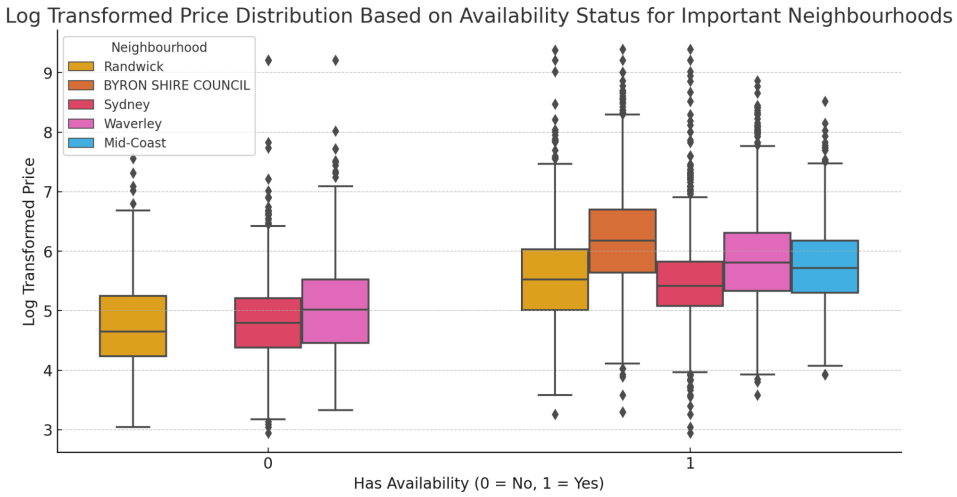


Figure A4. Price Distribution of Available Vs Unavailable Airbnb in Popular Neighbourhoods

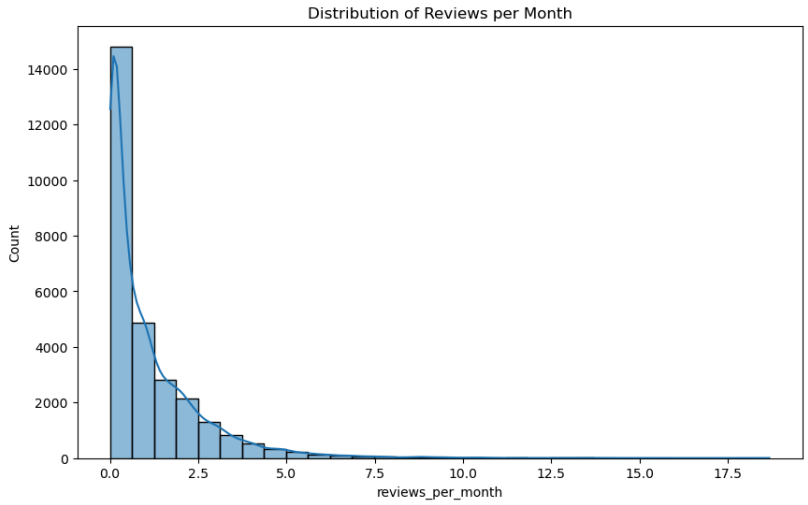
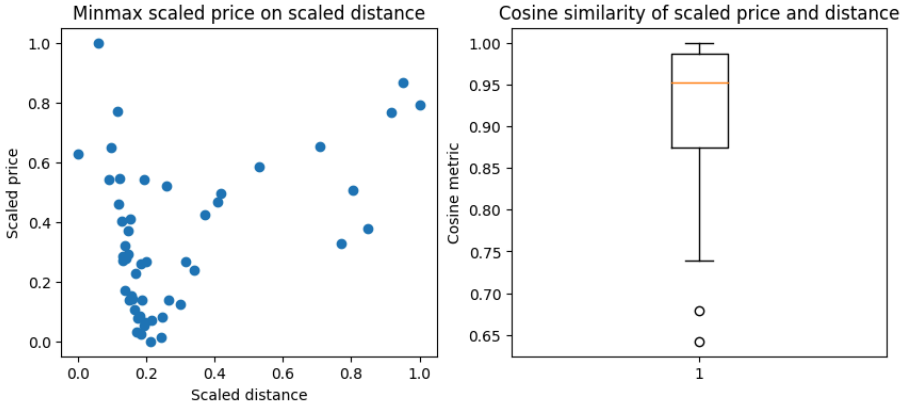


Figure A5. Number of Reviews



Appendix A6: Cosine similarity of price and distance

Appendix B: Methodology (Modelling)

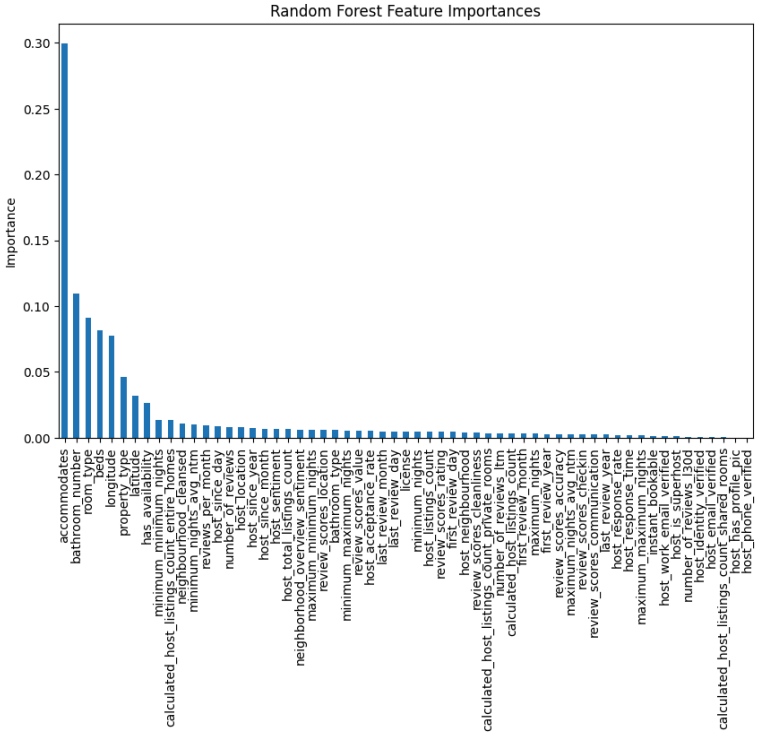


Figure B1. Random Forest Feature Importance

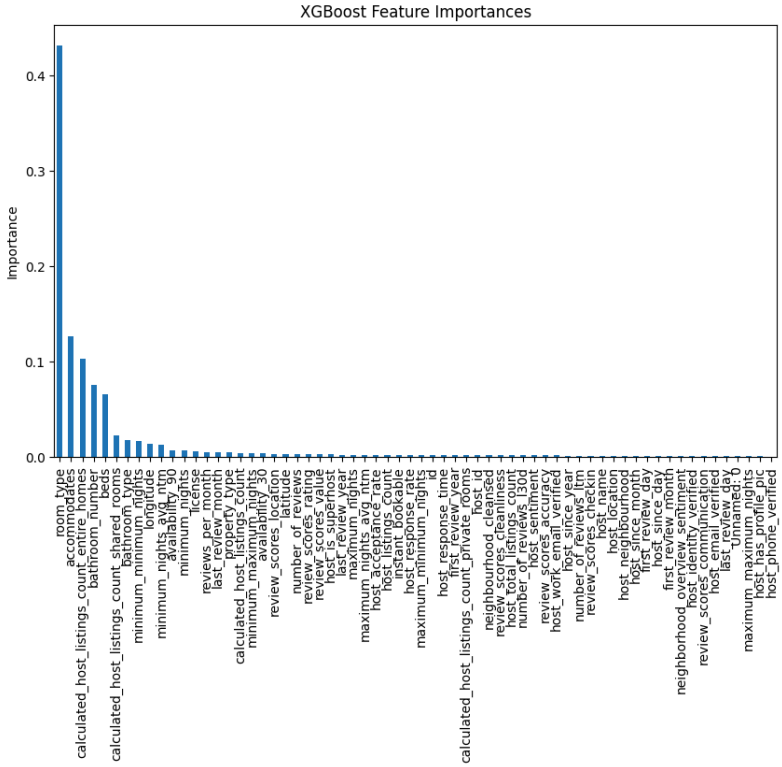


Figure B2. Feature Importance of XGBoost

Appendix C: Data mining: What are the best hosts doing?

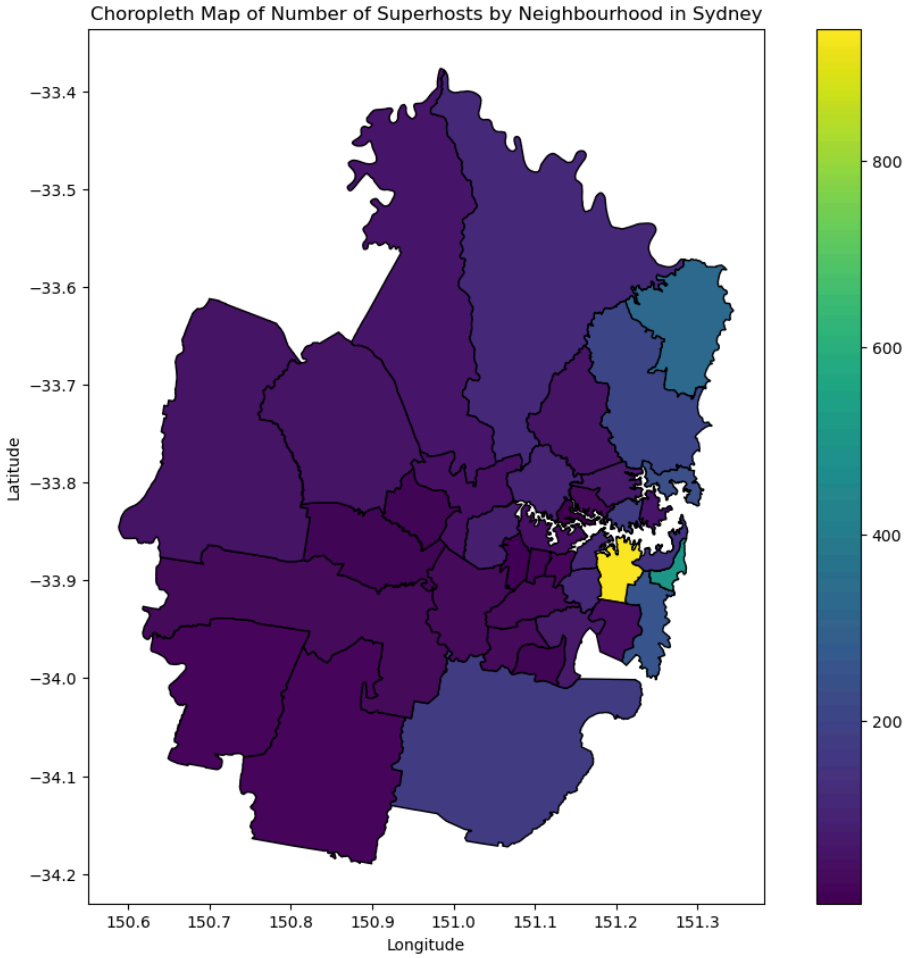


Figure C1. Number of Superhosts for Each Neighbourhood

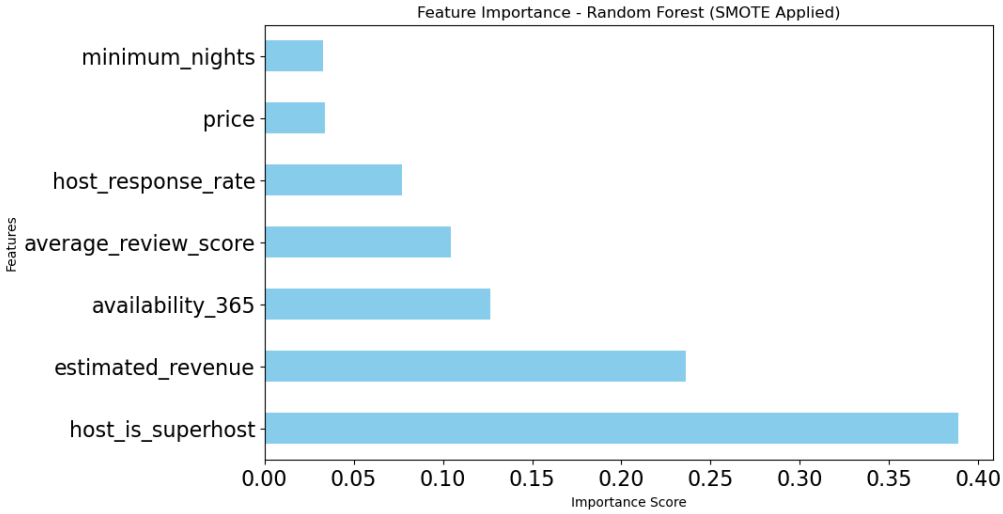


Figure C2. Feature Importance of Best Host Model

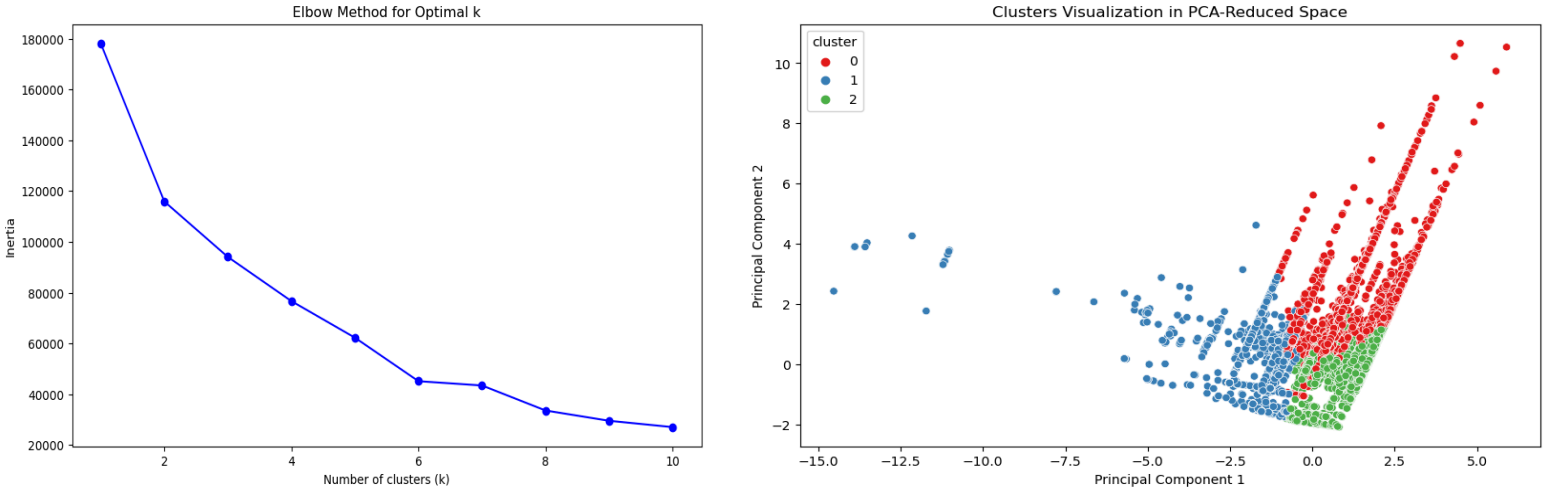


Figure C3. K-Means Clustering of Best Host Attributes

	precision	recall	f1-score
0	0.99	0.98	0.98
1	0.58	0.72	0.64
accuracy			0.97
macro avg	0.78	0.85	0.81
weighted avg	0.97	0.97	0.97

Table C1. Random Forest Classifier Evaluation Metrics

Host	Price Category	Availability Category	Review Category	Nights Category	Response Category	Superhost Category
Host 1	High Price	High Availability	High Review	Short Stay	Fast Response	Superhost
Host 2	High Price	Medium Availability	High Review	Short Stay	Fast Response	Superhost
Host 3	High Price	Medium Availability	Medium Review	Short Stay	Fast Response	Superhost
Host 4	High Price	High Availability	Medium Review	Short Stay	Fast Response	Superhost
Host 5	Medium Price	Medium Availability	High Review	Short Stay	Fast Response	Superhost

Table C2. A few example sequences for GSP

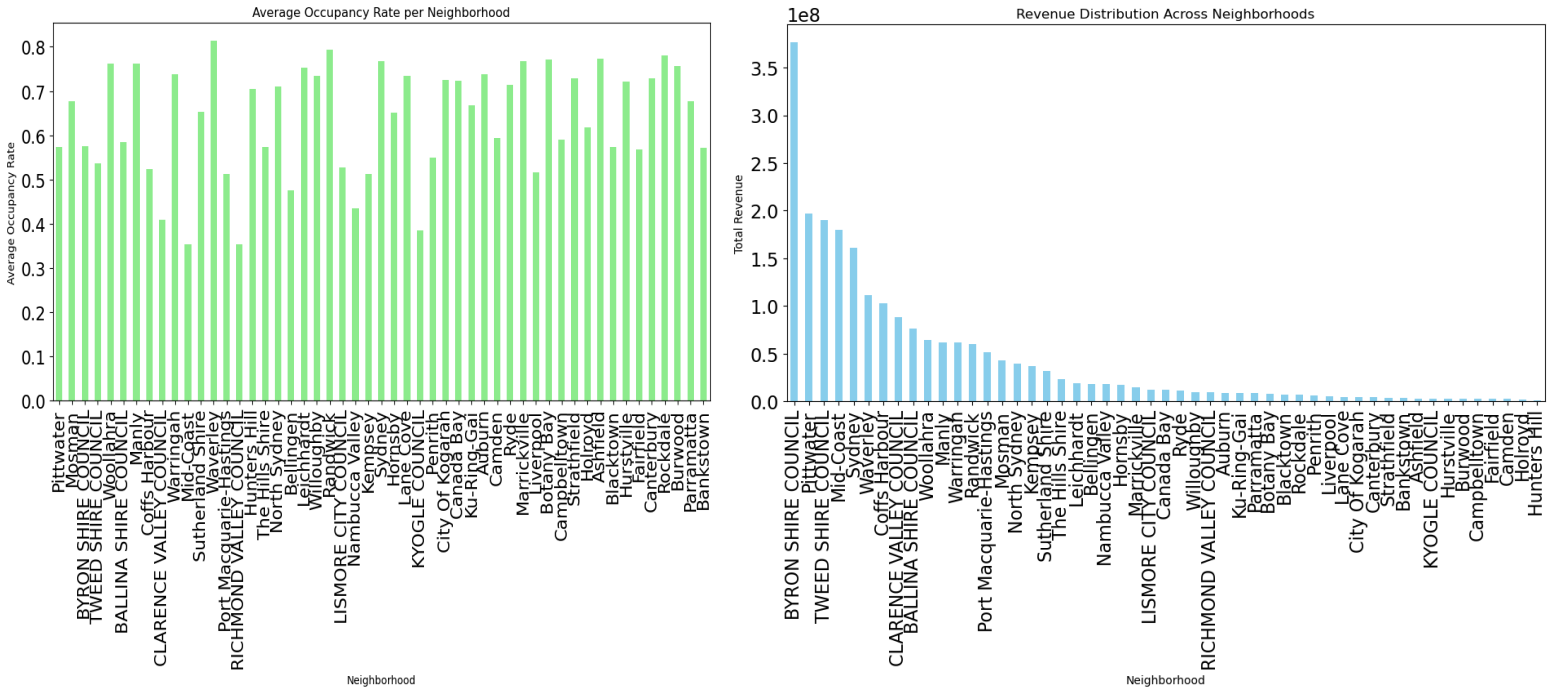


Figure C4. Revenue and Occupancy Rate for Neighbourhood

	price	average_review_score	revenue	availability_365	host_response_rate
count	1521	1521	1521	1521	1521
mean	896.07	4.9	170159	183.79	97.45
std	499.66	0.07	172476.9	130.35	0
min	497	4.7	0	0	100
25th Percentile (Q1)	590	4.86	35308	45	100
Median (Q2)	739	4.91	151216	209	100
75th Percentile (Q3)	1000	4.95	235900	313	100
max	6000	5	1602000	365	100

Table C3. Best Hosts Metrics

	price	average_review_score	revenue	availability_365	host_response_rate
count	34109.00	34109.00	34109.00	34109.00	34109.00
mean	358.97	3.58	57146.88	127.58	63.03
std	799.69	2.07	242241.10	131.83	47.11
min	17.00	0.00	0.00	0.00	0.00
25th Percentile (Q1)	130.00	2.71	0.00	0.00	0.00
Median (Q2)	220.00	4.76	16637.00	83.00	100.00
75th Percentile (Q3)	380.00	4.91	62150.00	256.00	100.00
max	100000.00	5.00	36500000.00	365.00	100.00

Table C4. Average Host Metrics

Superhost	price	revenue	occupancy_rate	average_review_score	host_response_rate
Non	364.96	57447.82	0.66	3.36	54.15
Superhost	433.86	76431.79	0.58	4.51	97.34

Table C5. Attribute metrics for Superhosts vs Non-Superhosts