Data processing

The training dataset contains highly detailed information on the 10,635 existing Airbnb listings in Sydney, and the test dataset includes information on 24,818 listings. Both datasets shared the same features. The training dataset was used to find the best predictive model, while the test dataset was used for model evaluation. The data within these datasets were cleaned, and missing values were addressed prior to exploratory data analysis (EDA).

Data cleaning

For computational convenience, the ‘id’ column was used as the index for the training and the test datasets. These two datasets contained the same 82 features that could potentially affect the prices, including text, numerical and categorical features. Additionally, the training dataset contains the response variable ‘price’. As both datasets required data processing, they were first combined for data cleaning and imputation and separated in later sections.

The features ‘host_response_rate’ and ‘host_acceptance_rate’ were converted to type ‘float’ as they were misclassified as objects. The dabl package was then used to detect column types and remove useless variables automatically. Seven variables were considered useless by the dabl package: ‘experiences_offered’, ‘bed_type’, ‘requires_license’, ‘is_business_travel_ready’, ‘require_guest_profile_picture’, ‘require_guest_phone_verification’ and ‘calculated_host_listings_count_shared_rooms’. They were considered useless because a significant proportion of their values were concentrated within one category, failing to provide representative information. Thus, they were removed to enhance the performance of the model. Furthermore, the variables ‘host_id’, ‘host_since’, ‘first_review’, and ‘last_review’ were examined and deemed irrelevant to the response variable based on the correlation analysis and domain knowledge.

Dealing with missing values

The proportion of missing values in each feature were calculated and displayed in descending order. Table 1 shows 32 features with null values, of which ‘square_feet’ has the most null values (35,279). Three variables (‘square_feet’, ‘weekly_discount’, ‘monthly_discount’) have more than 90% of missing values and were removed from the dataset.

Table 1: Missing Percentages

Predictor	Missing Percentage (%)	Predictor	Missing Percentage (%)	Predictor	Missing Percentage (%)
square_feet	99.509	transit	33.289	space	26.542
monthly_discount	95.290	host_acceptance_rate	32.319	reviews_per_month	24.723
weekly_discount	92.147	security_deposit_perc	32.088	cleaning_fee_perc	23.532
notes	57.936	host_neighbourhood	29.656	neighbourhood	12.792
host_about	45.170	review_scores_value	27.710	beds	0.669
access	42.682	review_scores_checkin	27.699	zipcode	0.324
host_response_rate	42.016	review_scores_location	27.693	host_location	0.107
host_resposn	42.016	review_score	27.662	city	0.087

e_time		**s_accuracy**			
house_rules	41.418	**review_scores_communication**	27.628	**bedrooms**	0.056
interaction	38.544	**review_scores_cleanliness**	27.617	**bathrooms**	0.025
neighborhood_overview	34.347	**review_scores_rating**	27.563		

Several approaches were considered when processing the null values. Firstly, the potential relationship between ‘host_resposne_time’, ‘host_response_rate’ and ‘host_acceptance_rate’ was explored. It was considered unreasonable to assume that over 40% of hosts did not respond to inquiries, as Airbnb reduces the search priority of listings with unresponsive hosts (Airbnb, n.d.). Given this assessment, the missing entries in the ‘host_response_rate’ and ‘host_acceptance_rates’ were replaced with their respective means. The missing values within ‘host_response_time’ were replaced with a new category of 'None', which was later merged with the category with the most similar statistical characteristics.

Secondly, it was determined based on domain knowledge that the observations for the 'review_score_location' feature were likely to be similar for properties located in the same 'street'. Hence, the missing values in this feature were filled with the mean value of other properties on the same street. The remaining empty values within this feature were replaced with the mean value of all observations. The missing observations in the ‘bedrooms’ and ‘beds’ features were replaced with median values based on the number of guests that the property could accommodate. The median was chosen to keep all observations within the feature in integer form. The missing values in the ‘bathrooms’ feature were replaced with median values according to the number of ‘bedrooms’ within the property. The missing values in the ‘bathrooms’ feature were replaced with median values according to the number of ‘bedrooms’ included in the listing. For the remaining features, the median observation of the feature was used to replace the null values in each numerical feature, and missing observations in other text features were replaced with ‘None’.

Subsequently, the dataset was reassessed to ensure all missing values had been handled. Of the ten features related to the listing location, the feature 'neighbourhood_cleansed' was selected to represent the location of the property as it contained the least errors with no missing values. Other features were subsequently discarded from the dataset to prevent issues with multicollinearity.