Exploratory Data Analysis (EDA)

Response variable: Price

Prior to conducting EDA, the training and test datasets were split, as EDA is performed only on the training dataset by convention. Following this split, the training dataset contained 10,635 observations. The maximum value of price was $4300.000, while the median price was $144.000, and the mean price was $212.000. The distribution plots of price and the descriptive statistics suggest that the response variable is highly right-skewed, with a skewness of 5.703 and kurtosis of 52.601, significantly deviating from a normal distribution and potentially indicating the presence of outliers. This deviation suggests that log transformation may be needed for linear models. Following a log transformation, the price becomes approximately normally distributed with a slight right skew, with skewness 0.850 and kurtosis 0.826.

Figure 1: Distribution Plot of Price

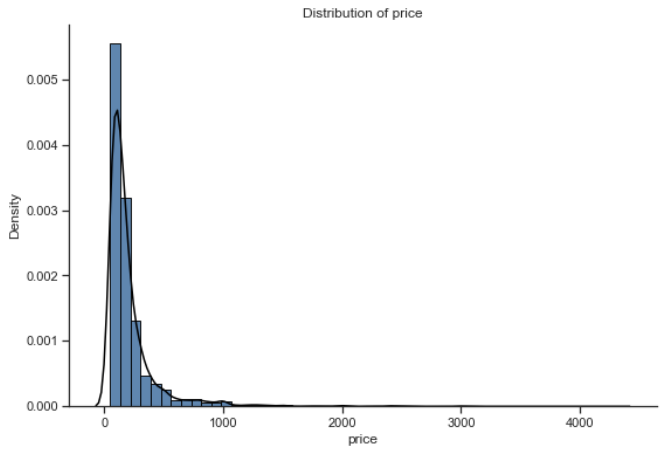


Table 1: Descriptive Statistics of Price

	Price
Count	10635.000
Mean	212.368
STD	255.625
Min	51.000
25%	89.000
50%	144.000
75%	231.000
Max	4300.000

Figure 2: Distribution Plot of Log Price

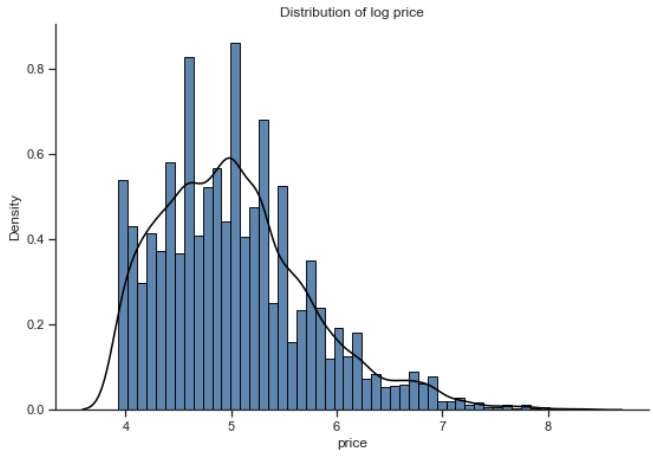


Table 3: Descriptive Statistics of Price

	Price
Count	10635.000
Mean	5.035
STD	0.723
Min	3.932
25%	4.489
50%	4.970
75%	5.442
Max	8.366

Numerical variables

A correlation heatmap was created to explore the bivariate relationships between numerical variables. The heatmap (Appendix A) indicates multiple highly correlated features that need to be deleted to reduce multicollinearity. To assess which features needed to be discarded, further assessment was conducted through the construction of a correlation and mutual information table to evaluate the relationship between the predictors, as well as the relationship between the predictors and the response variable.

1. As shown in Table 4, four features related to the total listings of each host showed a correlation above 0.9, raising significant concerns regarding multicollinearity between these features. Despite this, all four features demonstrated a correlation below 0.2 with the response variable, indicating a weak linear relationship with price. Thus, from the listed features, only 'calculated_host_count_entire_homes' will be kept in the data set as it has the most significant linear relationship with price, with a correlation of 0.361.

Table 4: Correlation Table of Host Listings

	host_listings_count	**host_total_listings_count**	**cal_host_listings_count**	**cal_host_count_entire_homes**
host_listings_count	1.000			
host_total_listings_count	1.000	1.000		
cal_host_listings_count	0.937	0.937	1.000	
cal_host_count_entire_homes	0.932	0.932	0.994	1.000
Price CORR	0.164	0.166	0.164	0.361
Price MI	0.182	0.182	0.168	0.181

2. Eight features were related to the minimum and maximum nights that a guest could stay. Among these eight features, 'min_average_nights_ntm' and 'max_average_nights_ntm' will be kept for further data analysis based on their low correlations with each other (Table 5), and their relatively high mutual information and correlation with price.

Table 5 Correlation Table of Max & Min Nights

	Min	**Max**	**Min_min**	**Max_min**	**Min_max**	**Max_max**	**Min_avg**	**Max_avg**
Min	1.000							
Max	0.003	1.000						
Min_min	0.863	-0.012	1.000					
Max_min	0.868	-0.002	0.957	1.000				
Min_max	-0.002	0.008	-0.002	-0.001	1.000	1.000		
Max_max	-0.002	0.008	-0.002	-0.001	1.000	1.000		
Min_avg	0.880	-0.003	0.978	0.992	-0.002	-0.002	1.000	
Max_avg	-0.002	0.008	-0.002	-0.001	1.000	1.000	-0.002	1.000
Price CORR	0.012	-0.015	0.015	0.024	0.010	0.010	0.017	0.010
Price MI	0.0889	0.108	0.0696	0.0971	0.0452	0.0289	0.114	0.0337

3. Among variables related to the accommodation of guests, the feature 'accommodates' will be removed as its correlation with both bedrooms and beds is greater than 0.8

(Table 5).

Table 6: Correlation Table of Accommodates, Bathrooms, Bedrooms & Beds

	accommodates	bathrooms	bedrooms	beds
accommodates	1.000			
bathrooms	0.549	1.000		
bedrooms	0.842	0.598	1.000	
beds	0.871	0.520	0.808	1.000
Price CORR	0.574	0.497	0.600	0.519
Price MI	0.360	0.227	0.365	0.263

4. Features regarding availability suggest the available days of a listing for the next 30/60/90 days and one year. Among these four integer variables, 'availability_30' is kept for further data analysis as it has the highest mutual information with price (Table 7).

Table 7: Correlation Table of Availability

Availability	_30	_60	_90	_365
_30	1.000			
_60	0.947	1.000		
_90	0.893	0.977	1.000	
_365	0.632	0.704	0.747	1.000
Price CORR	0.094	0.089	0.087	0.124
Price MI	0.044	0.043	0.040	0.034

5. There are seven features related to the review score: rating, accuracy, cleanliness, check-in, communication, location and value. Among these features, 'review_score_rating' will be deleted as it is an overall rating created by Airbnb with its algorithm based on other scores.

Table 8: Correlation Table of Review Scores

	Rating	Accuracy	Cleanliness	Checkin	Communication	Location	Value
Rating	1.000						
Accuracy	0.752	1.000					
Clean	0.722	0.660	1.000				
Checkin	0.581	0.555	0.457	1.000			
Commun	0.642	0.619	0.488	0.693	1.000		
Location	0.486	0.484	0.412	0.432	0.466	1.000	
Value	0.719	0.689	0.656	0.521	0.569	0.516	1.000
Price Corr	0.058	0.039	0.049	0.040	0.031	0.056	-0.007
Price MI	0.034	0.073	0.045	0.085	0.113	0.113	0.058

6. Among the features related to the number of reviews, the number of reviews for the property over the most recent twelve months, i.e., 'number_of_reviews_ltm', has been discarded as it had a significant correlation with 'reviews_per_month' (Table 9).

Table 9: Correlation Table of Reviews

	number_of_reviews	number_of_reviews_ltm	reviews_per_month

number_of_reviews	1.000		
number_of_reviews_ltm	0.784	1.000	
reviews_per_month	0.665	0.858	1.000
Price Corr	-0.072	-0.071	-0.066
Price MI	0.034	0.039	0.058

Categorical variables

There are 9 categorical variables in the training dataset. The boxplot (Appendix B), table 10 and table 11 show that 'host_is_superhost', 'host_identity_verified', 'is_location_exact' and 'instant_bookable', have similar median and mean listing prices among the different categories. Therefore, these four variables will be discarded as they are not representative.

Table 10: Median Price

Median price ($)	False	True
host_is_superhost	142.000	149.000
host_identity_verified	140.000	149.000
is_location_exact	140.000	144.000
instant_bookable	150.000	138.000

Table 11: Mean Price

Mean Price ($)	False	True
host_is_superhost	211.842	215.394
host_identity_verified	211.358	214.191
is_location_exact	204.992	214.699
instant_bookable	230.598	189.610

Detailed EDA on remaining variables

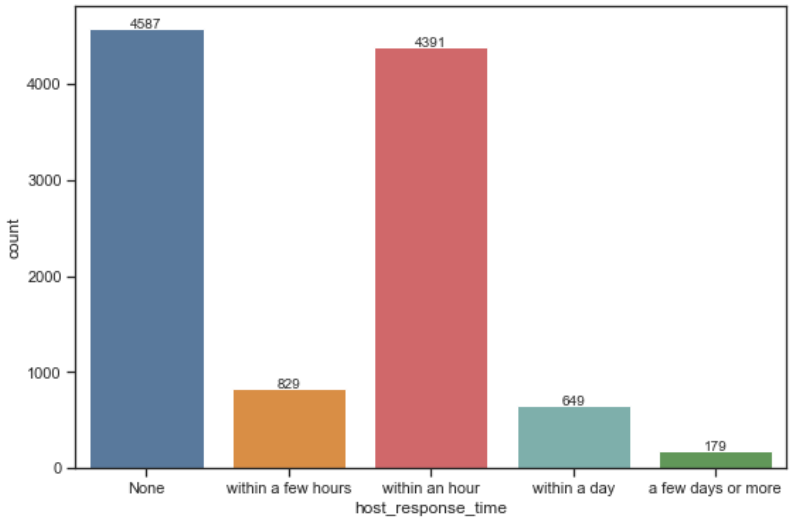
Host_response_rate & host_acceptance_rate: The missing values in these two features are imputed with their respective means to preserve information, as the vast majority of the observations within these two features are concentrated between 0.9 and 1. However, they appear to have a weak relationship with price (Appendix C). As the missing values within these two features were replaced with their respective mean values, both features demonstrate a significantly left-skewed distribution.

Host_response_time: This feature contains five different categories. Of these categories, Table 12 shows that 'within a few hours' and 'within an hour' had the highest median value of price while 'a few days or more' had the lowest median price. Figure 3 shows that most observations are concentrated in 'None' and 'within an hour'.

Table 12: Descriptive Statistics of host_response_time

	None	A few days or more	Within a day	Within a few hours	Within an hour
Count	4587	179	649	829	4391
Median	129	110	140	150	150

Figure 3: Count Plot of host_response_time



Neighborhood_cleansed: The count plot (Appendix E) shows that Sydney (2890 observations), Waverley (1469) and Randwick (890) have the most listings. This outcome is unsurprising, as Sydney is the most popular place for travellers in Australia and Waverley & Randwick have famous tourist attractions. Among all neighbourhoods, Pittwater has the highest median price at $350, followed by Manly ($195), Hunters Hill ($194.5), Mosman ($185) and Waverley ($165). These five neighbourhoods are located near beaches and famous tourist attractions and are thus likely to be in greater demand, which could be associated with higher prices, especially around peak tourist season.

Table 13: Median Prices of Different Neighbourhoods

Price ($)	Highest 5	Price ($)	Lowest 5
Pittwater	350.00	**Campbelltown**	89.00
Manly	195.00	**City of Kogarah**	88.50
Hunters Hill	194.50	**The Hills Shire**	85.00
Mosman	185.00	**Blacktown**	79.00
Waverley	165.00	**Holroyd**	74.00

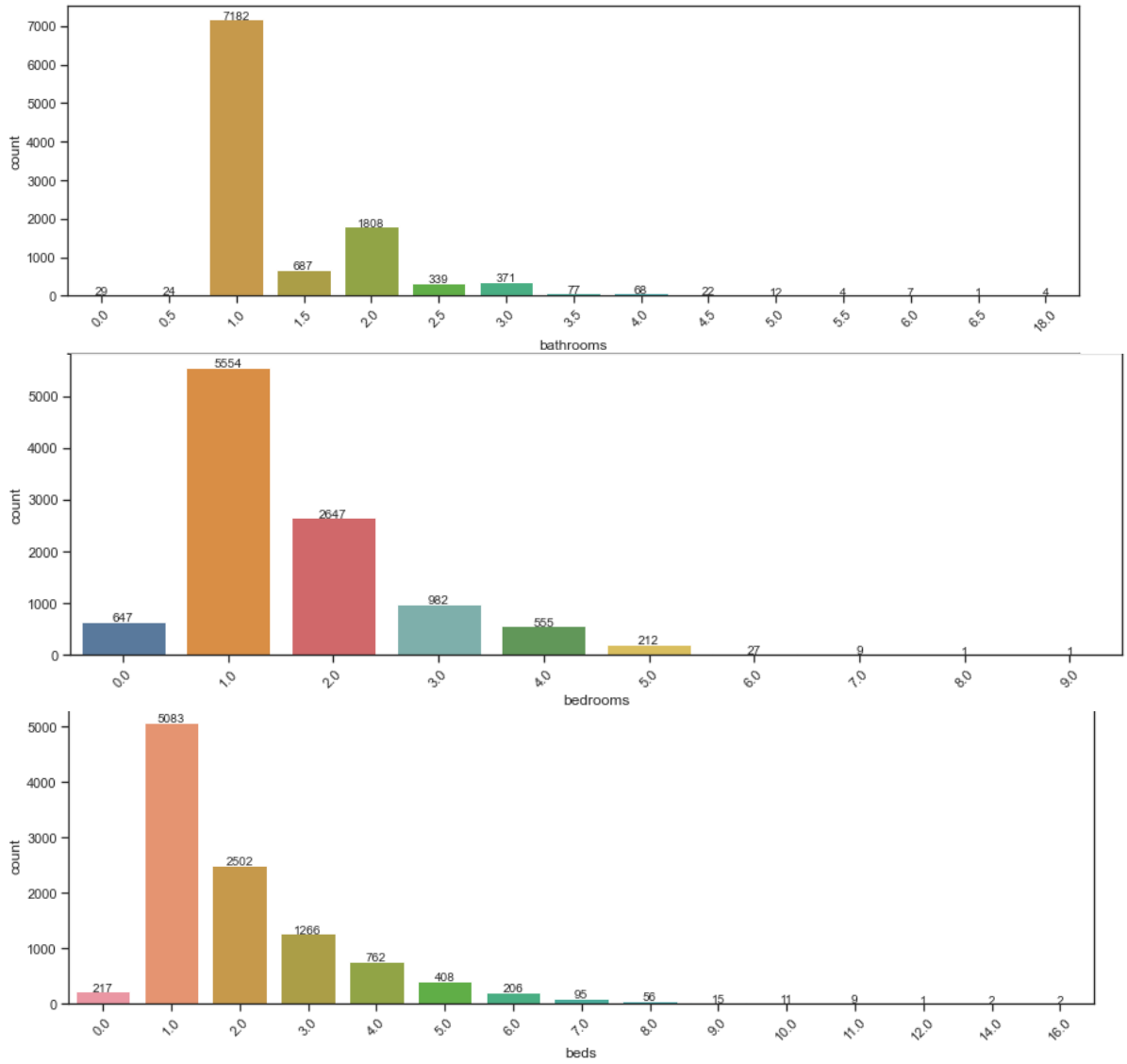
Property_type: Most observations within this feature were concentrated in 'Apartment' (6492 observations) and 'House' (2671 observations) (Appendix F). Among the different property types, boat and castle had the highest median price. However, they are sparse categories, with only 10 and 1 observations, respectively.

Room_type: (Appendix F) there are four different types of room; of these, hotel rooms had the highest median price ($191) and shared room had the lowest ($70). Most observations were concentrated in 'Entire home/apt' (7202 observations) and 'Private room' (3307 observations).

Cancellation_policy: There were seven categories within this feature, among which, two luxury cancellation policies, ‘luxury_moderate’ and ‘luxury_super_strict_125’, had the highest median price. The flexible cancellation policy had the lowest median price ($110) (Appendix F). These findings are intuitive as the cancellation policy refers to the lead time allowed for free cancellation. As the policy becomes stricter, the cancellation fee increases to protect the interests of the hosts. The observations are mostly concentrated in ‘flexible’ (3312 observations), ‘moderate’ (2586 observations), and ‘strict_14_with_grace_period’ (4651 observations).

Bathrooms, Bedrooms & Beds: Figure 3 shows that the observations for these features were concentrated at specific values. In the case of bathrooms, the observations are concentrated at 1.0 and 2.0. For bedrooms, the observations were concentrated between 1.0 to 3.0. For the number of beds, the observations were concentrated between 1.0 to 4.0. The boxplots suggest that these three features are ordinal, as the price of properties increased with the number of bathrooms, bedrooms and beds (Appendix G).

Figure 4: Count Plot of Bathrooms, Bedrooms & Beds



Security_deposit_perc, cleaning_fee_perc & extra_people_perc: The distributions of these features demonstrated a significant positive skew, indicating that most hosts keep the deposit, cleaning fee and the fee for additional guests within a certain range unless the room or house met other specific criteria (Appendix H). Transformations such as the log, box-cox and Yeo-Johnson transformation may be needed. Analysis of the features also revealed the presence of outliers. However, these outliers were not removed because there is no information suggesting that these observations were illegitimate.

Guest_included: The observations for this feature ranged from 1 to 15, concentrating on 1.

The number of guests included was found to positively correlate with the response variable, indicating that this feature should be considered an ordinal variable (Appendix I).

Availability_30: This feature represents the availability of a property for the next 30 days as set by the host. The observations ranged from 0 to 30, concentrating on 0 (6035 observations) (Appendix I).

Reviews: Although 'number_of_reviews' and 'reviews_per_month' are highly correlated with a correlation of 0.665, they have both been retained as part of the training set, given that they convey different messages. Transformations are required since they are highly right-skewed (see Appendix J). Transformations are required since they are highly right-skewed (Appendix J).

Review scores: there are six different review scores, including accuracy, cleanliness, check-in, communication, location, and value. These features are nominal variables that act as a good indicator of customer satisfaction. The distribution plots (Appendix K) suggests that most customers' aggregate review scores were higher than 0.8. The regression plots suggest that a property's review score is positively correlated with its price, indicating that this feature should be considered an ordinal variable (Appendix K).

Calculated_host_listings_count_entire_homes: This feature represents a host's total listings for entire homes, while **calculated_host_listings_count_private_rooms** represents the number of a host's total listings for private rooms. These two features are highly right-skewed with a concentration from 0 to 2. These two features are highly right-skewed, and observations are concentrated between 0 to 2 (Appendix L).