Airbnb: The Sydney Story

Junjie Qiu[1] **Wang MaQiangHu**[1] **Qiang Hu**[1] **Xinyi Zhu**[1]

[1]STATDS, SUSTech, Shenzhen, China

[12111831,12012424,12111214,12111214,12112944]@mail.sustech.edu.cn

Abstract

Our study provides an exploratory data analysis and a predictive model of Airbnb listings in Sydney to gather insights for potential travelers and hosts. The analysis encompasses neighborhood safety, listing distribution, and pricing patterns. Additionally, clustering analysis identifies distinct groups of listings based on property characteristics and their relation to pricing. A linear regression model, enhanced by variable selection and transformation, predicts listing prices with substantial accuracy. The findings highlight the importance of neighborhood safety and property features in determining accommodation choices and pricing strategies on Airbnb.

1 Exploratory Data Analysis

Airbnb is one of the largest companies offering hotel-like short-term rentals. As our group is passionate about traveling, we decided to analyze the Airbnb market in Sydney to gather insights before starting our journey to Australia. Here, we conduct some exploratory data analysis to begin with.

1.1 Neighborhoods

1.1.1 Listing count per neighborhood

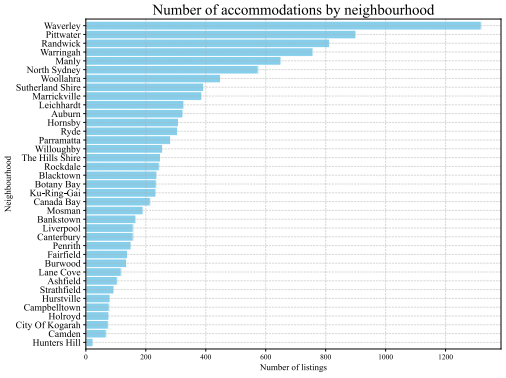

Figure 1: Number of listings by neighborhood

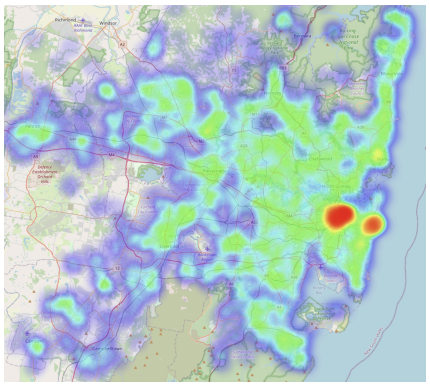

Figure 2: Heat map of distribution of accommodations

Figure 1 shows that neighborhood "Manly" holds most listing, and altogether 20 neighborhoods have over two hundred accommodations. In Figure 2, we can see that most of listings are in the city's bay area. In addition, Figure 2 can better show to which extent the accommodation clustered than Figure 1.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

1.1.2 Average price per neighborhood

To ensure a fair comparison, only the most common type of accommodation, which is for two persons, is selected. As anticipated, the priciest accommodations are predominantly situated in the city's bay area.

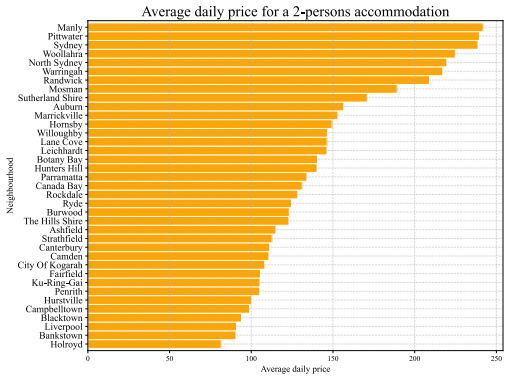

Figure 3: Average daily price for a 2-persons accommodation

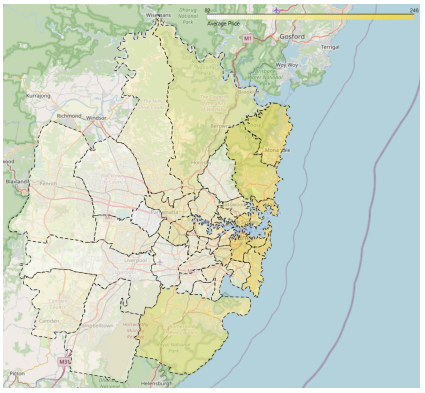

Figure 4: Average daily price map for a 2-persons accommodation

1.1.3 Neighborhood safety

Figure 5 shows the number of homicides, robberies, and thefts per neighborhood over two years. It is evident that theft constitutes the majority of crimes in Sydney. To obtain a comprehensive measure of overall safety, we combined the counts of these three types of crimes by summing their standardized scores to create a crime index. Figure 6 presents the crime index across neighborhoods. Based on the findings from Section 1.1.2, we recommend that tourists select neighborhoods where accommodations are reasonably priced and the area is safe.

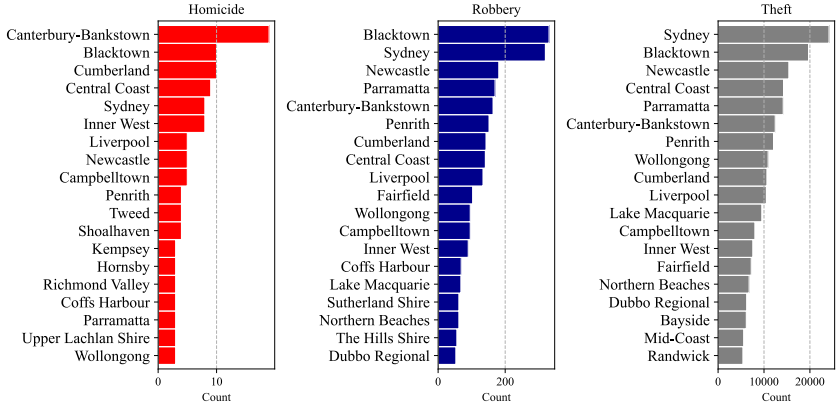

Figure 5: The number of homicide, robbery, and theft per neighborhoods in two years

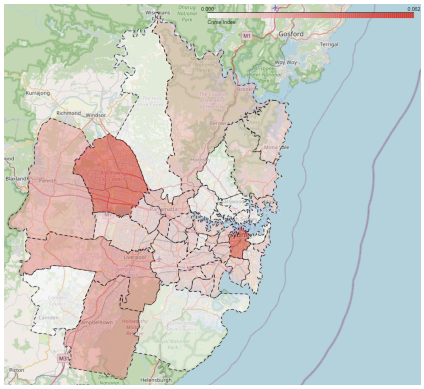

Figure 6: Overall crime index

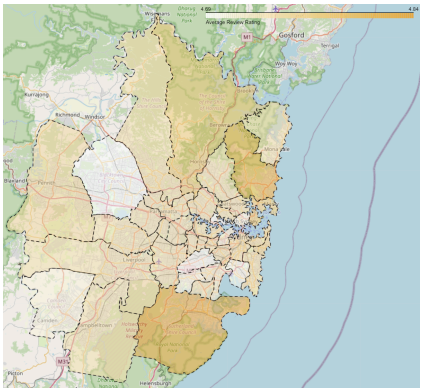

Figure 7: Average review rating

1.2 Review score

Figure 7 shows the average review rating per neighborhood. When compared with Figure 6, it is evident that areas with a high crime index generally have lower review ratings. We recommend Sutherland Shire to tourists, as it is a safe, reasonably priced neighborhood with high review ratings.

1.3 Room type and property type

From Figure 8, we observe that entire homes/apartments are the most popular type of accommodation on Airbnb in Sydney, accounting for more than two-thirds of the market. Figure 9 provides the price per accommodation for each room type. In terms of cost performance, we can conclude that

```
Private room > Shared room > Entire home/apartmen > Hotel room
```

, which indicates that compared to other types of accommodations, hotels are not an economical choice on Airbnb.

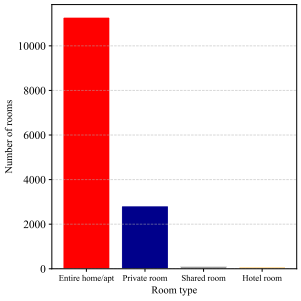

Figure 8: Number of accommodations per room type

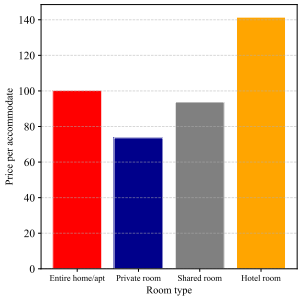

Figure 9: Price per accommodations per room type

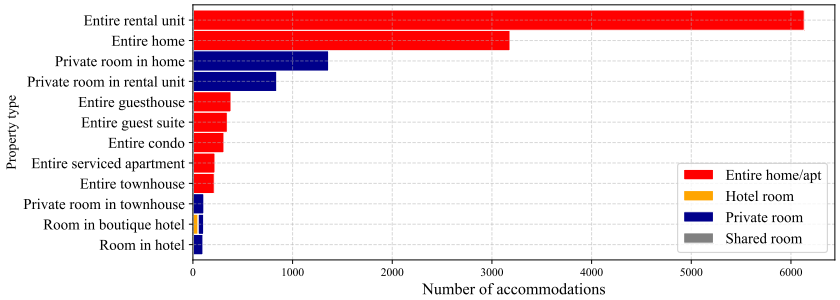

Figure 10: Number of accommodations per property type

Figure 10 displays the number of accommodations per property type. The most frequent property type is `Entire rental unit`. Some accommodations with the property type `Room in boutique hotel` or `Room in hotel` are labeled as `Private room`. It is unclear whether this phenomenon is due to manual misclassification.

1.4 Average price by date

Figure 11 shows the average price of available accommodations by date, from March 2024 to March 2025, specifically for accommodations for two persons. The data was compiled on March 16th, 2024, leading to the suspicion that prices for dates further in the future have not been updated yet and are likely default prices. For instance, there is a noticeable dramatic decline in prices on June 15th, 2024. Additionally, it appears that the availability of accommodations is predominantly set for one season, resulting in a periodic, stair-step-like decline in the number of available accommodations.

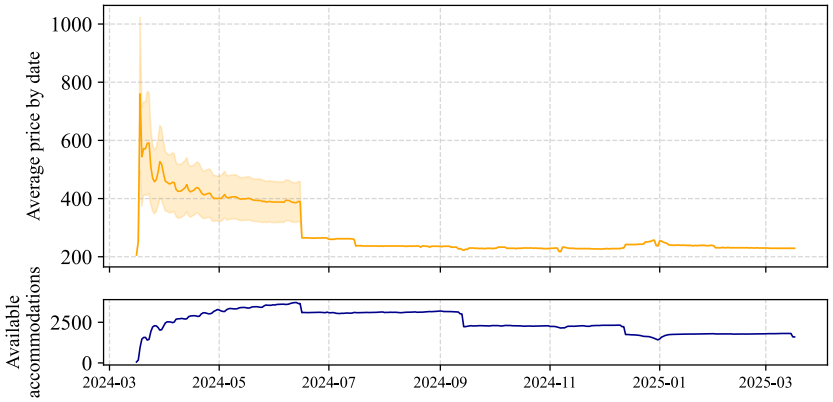

Figure 11: Average price and number of available accommodations by date

2 Cluster Analysis

This part details a clustering analysis performed on Airbnb listings in Sydney, utilizing key property characteristics to group similar properties. The primary objective is to determine the relationship between property features per capita and their influence on pricing. Using the K-Means clustering algorithm and the Elbow method for model evaluation, the analysis identifies distinct property clusters with varying price levels. The findings suggest nuanced relationships between property features and pricing, providing insights for both hosts and potential guests.

2.1 Relevant columns for clustering

The following features describe the type and size of the listing, which are important for clustering similar properties together:

- property_type
- room_type
- accommodates
- bedrooms
- bathrooms
- beds
- price.

Since for the categorical features like room_type, the numbers of different categories vary a lot, say, they are unfair, if we directly use these unfair categories, such as, the percent of “entire room” type is larger than 50%. These unfair categories will destroy the clustering result, so we will not use the categorical data in this part.

Therefore, we will use the following feature later: accommodates

- bedrooms

- bathrooms
- beds
- price.

2.2 Data transformation

To normalize the property characteristics and account for the varying capacities of listings, the following per capita metrics were calculated:

- Per_person_bedrooms: Bedrooms divided by accommodates
- Per_person_bathrooms: Bathrooms divided by accommodates
- Per_person_beds: Beds divided by accommodates

These transformations ensure that the clustering process is not biased by the absolute size of the properties but rather reflects the relative availability of amenities.

Another concern pushed us to make this transformation is that the absolute value of accommodates, bedrooms, bathrooms and beds can partially reveal the room_type, which is a unfair category as we discussed before. Since, for a "entire room" type listing, it normally has a larger accommodates, bedrooms and beds than a "private room" type listing published by a personal host. Thus, here we use a per capita metrics to escape from this kind of unfairness, and we will then to discover how the per capita metrics relates to the price of the listing.

2.3 Clustering algorithm

2.3.1 K-Means clustering

K-Means is a widely used clustering algorithm due to its simplicity and efficiency. It partitions the dataset into K distinct clusters, where each data point belongs to the cluster with the nearest mean. The algorithm involves the following steps:

1. Initialization: Randomly select K initial centroids.
2. Assignment: Assign each data point to the nearest centroid.
3. Update: Calculate the new centroids as the mean of all points in each cluster.
4. Repeat: Repeat the assignment and update steps until the centroids no longer change significantly.

Mathematically, the objective of K-Means is to minimize the within-cluster sum of squares (WCSS), defined as:

$$\text{WCSS} = \sum_{k=1}^{K} \sum_{x \in C_k} \parallel x - \mu_k \parallel^2 \tag{1}$$

where C_k is the k-th cluster, x is a data point, and mu_k is the centroid of C_k.

2.3.2 Elbow method

The Elbow method is used to determine the optimal number of clusters (K) for K-Means. It involves plotting the WCSS against the number of clusters and identifying the "elbow" point where the rate of decrease sharply slows down. This point indicates the number of clusters that best balance complexity and the goodness of fit.

Mathematically, the WCSS is computed for different values of K:

$$\text{WCSS}(K) = \sum_{k=1}^{K} \sum_{\{x \in C_k\}} \parallel x - \mu_k \parallel^2 \tag{2}$$

The Elbow point is identified where the first derivative of WCSS with respect to K shows a significant change in slope.

2.3.3 Evaluation metrics

Within-Cluster Sum of Squares (WCSS).

WCSS measures the compactness of the clusters. For each cluster, it calculates the sum of the squared distances between each point and the cluster centroid. A lower WCSS indicates more cohesive clusters. The goal of K-Means is to minimize WCSS across all clusters, ensuring that each cluster is as tight as possible.

Silhouette Score.

The Silhouette score is another metric used to evaluate the quality of the clusters. It measures how similar a data point is to its own cluster (cohesion) compared to other clusters (separation). The score ranges from −1 to 1, where a higher value indicates better-defined clusters. It is calculated as:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{3}$$

where $a(i)$ is the average intra-cluster distance and $b(i)$ is the average nearest-cluster distance for point i.

2.4 Clustering results

2.4.1 Determining the optimal number of clusters

Using the Elbow method, we plotted the WCSS against different values of K. The plot indicated an elbow at $K = 4$, suggesting that four clusters provide a good balance between minimizing WCSS and avoiding overfitting.

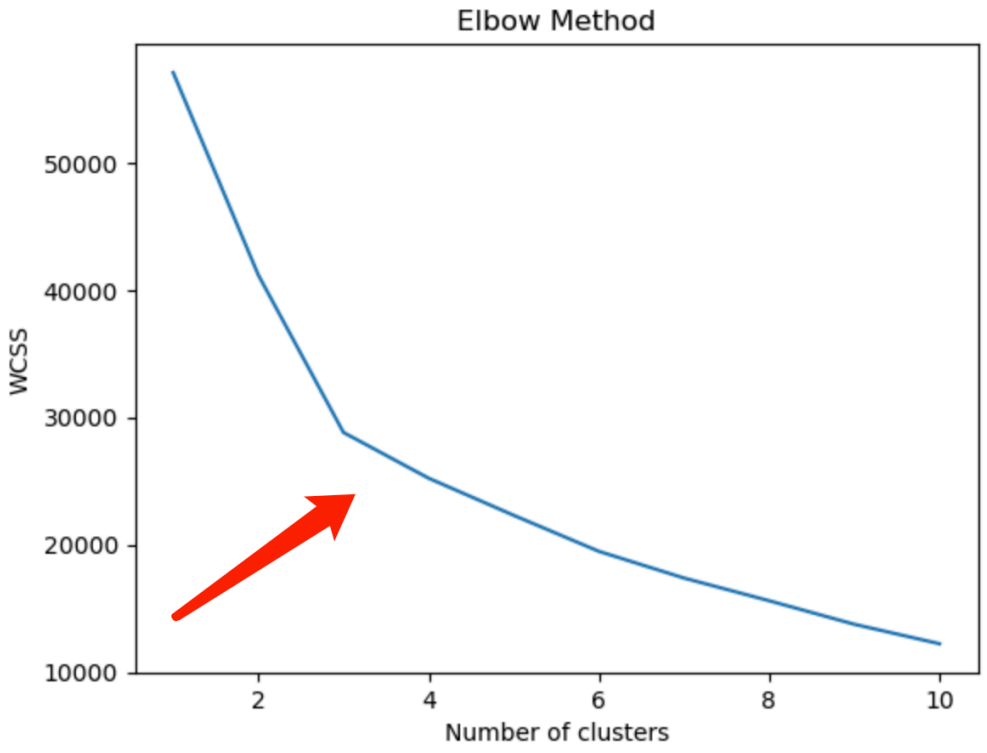

Figure 12: The choice of the number of clusters using Elbow methods.

2.4.2 Cluster characteristics

kmeans_cluster	per_person_bedrooms	per_person_bathrooms	per_person_beds	price	counts
0	1.091356	1.164754	1.047465	121.234099	977
1	0.436862	0.509594	0.442936	215.639481	6267
2	0.441414	0.296377	0.625936	383.478283	6349
3	0.496839	0.392673	0.584297	1949.097432	647

Figure 13: The clustering result table.

The analysis identified four distinct clusters based on the transformed features and their price levels:

- Cluster 0 (Low Price): Characterized by high bedrooms, high bathrooms, and high beds per capita.
- Cluster 1 (Mid Price): Characterized by moderate bedrooms, average bathrooms, and low beds per capita.
- Cluster 2 (High Price): Characterized by moderate bedrooms, low bathrooms, and moderately high beds per capita.
- Cluster 3 (Luxury): Represented by luxury holiday homes with high-end features.

2.5 Cluster analysis

A detailed examination of each cluster provided insights into the typical properties and possible explanations for their pricing:

2.5.1 Cluster 0 (Low Price)

- **Characteristics**: High per_person_bedrooms, per_person_bathrooms, and per_person_beds.
- **Example**: A property accommodating 4 persons with 5 bedrooms, 5 bathrooms, and 5 beds, yet priced low.
- **Assumptions**: This kind of rooms might mainly be **Private Rooms**, which means there are other peoples already living in the rooms (the owner of the house or other tenants). The existence of unknown roommates can be a cause of low price.
- **Room Type Distribution**:

room_type	Private room	Entire home/apt	Shared room	Hotel room
counts	817	102	50	8

2.5.2 Cluster 1 (Mid Price)

- Characteristics: Moderate per_person_bedrooms, average per_person_bathrooms, and low per_person_beds.
- Insight: Represents balanced properties suitable for mid-range budgets.

2.5.3 Cluster 2 (High Price)

- Characteristics: Moderate per_person_bedrooms, low per_person_bathrooms, and moderately high per_person_beds.
- Insight: Indicates properties with fewer bathrooms but more beds, justifying higher prices due to larger group accommodations.

2.5.4 Cluster 3 (Luxury)

- Characteristics: High-end properties with extensive amenities.
- Insight: Luxury holiday homes commanding premium prices due to superior features and exclusivity.

2.6 Visualization of clusters

We can simply distribute the four clusters in four quadrants in a plane, while

- x-axis represents the average bedrooms per capita
- y-axis represents the average bathrooms per capita.

Finally, we find the relationship: the average price increases with the quadrants.

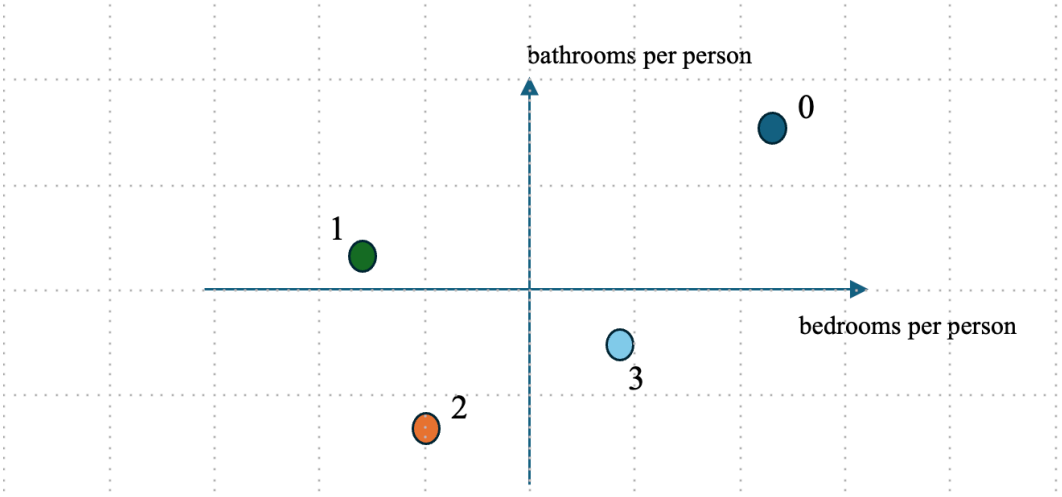

Figure 14: The illustration of per capita metrics and different clusters(with different prices)

2.6.1 Analysis and interpretation

The clustering results reveal that there is no direct linear relationship between the number of bedrooms, bathrooms, and beds per capita and the pricing of the listings. However, the clusters provide a framework to understand the diverse accommodation offerings and their market positions. The average price tends to increase with the higher quadrants on the per capita amenities plane, indicating a general trend where more exclusive features per capita correlate with higher prices.

2.7 Conclusion of clustering analysis

This clustering analysis of Airbnb listings in Sydney highlights the heterogeneity of the market and the complex interplay between property features and pricing. By transforming property characteristics to per capita metrics and employing the K-Means algorithm, meaningful clusters were identified, each representing distinct types of accommodations. The insights derived from this analysis can aid hosts in optimizing their pricing strategies and help guests in making informed decisions based on their accommodation preferences.

3 Prediction: Linear regression

3.1 Data processing

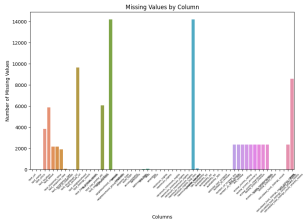

Figure 15: missing value for variables before variable selection

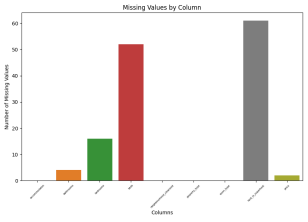

Figure 16: missing value for variables before variable selection

3.1.1 Missing value & outliers

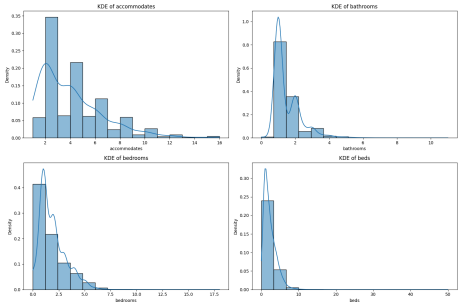

Figure 17: distribution of numerical variables before processing

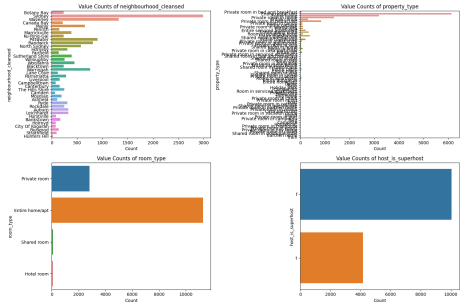

Figure 18: value counts of categorical variables before processing

- **Missing Values**: For independent variables, fill missing values (mode for categorical variables, mean for numerical variables). Remove instances with missing values in the response variable.
- **Outliers**: For numerical variables, handle outliers using the Interquartile Range (IQR) method or z-score(z-score leads to better R-squared and adjusted R-squared). For categorical variables, remove categories with too few instances (less than 10).

3.1.2 Data transformation

- Convert all categorical variables to dummy variables.
- Based on the distribution plots, the numerical variables deviate from a normal distribution. Apply the Box-Cox transformation to all numerical variables. (Also tried Yeo-Johnson transformation, the R^2 and adjusted R^2 are worse than Box-Cox).

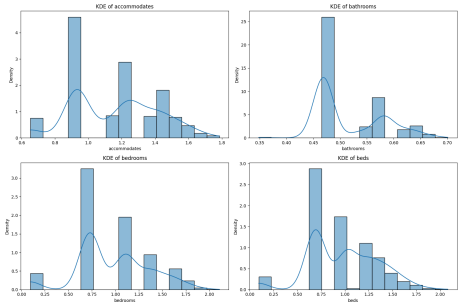

Figure 19: distribution of numerical variables after processing

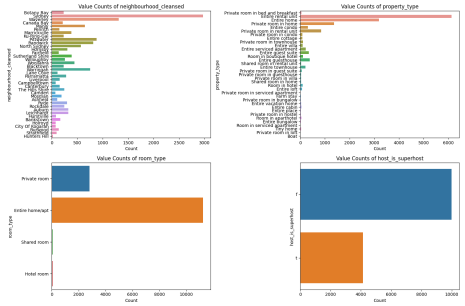

Figure 20: value counts of categorical variables after processing

3.2 Model selection

3.2.1 Variable selection

3.2.1.1 Basic manual selection

- **Correlation Ranking for Numerical Variables**

By sorting the correlation of numerical variables, select the top four as our feature variables: [‘accommodates’, ‘bathrooms’, ‘bedrooms’, ‘beds’].

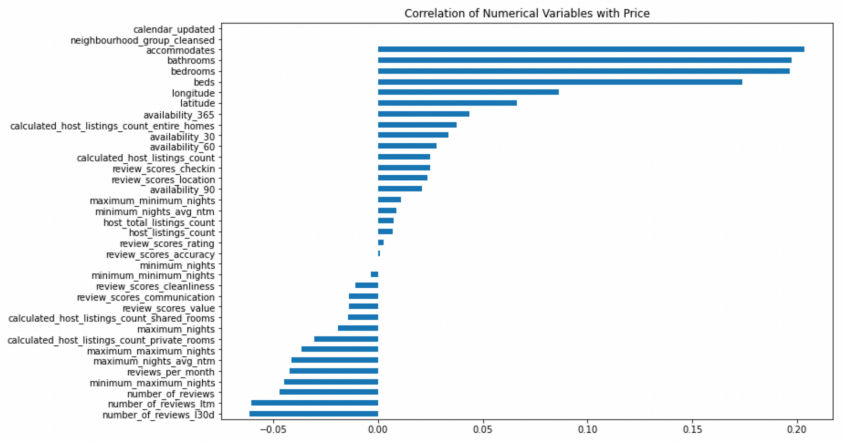

Figure 21: Correlation of Numerical Variables with Price

- **One Way ANOVA for Categorical Variables**

ANOVA results assist in feature selection by identifying which categorical variables should be included in the model based on their statistical significance. The results suggest that the categorical variables tested [‘neighborhood’, ‘property type’, ‘room type’] significantly influence the Airbnb rental prices in Sydney.

Table1. ANOVA Results for Categorical Variables

Index	ANOVA Neighbourhood Cleansed	ANOVA Property Type	ANOVA Room Type
C(neighbourhood_cleansed)	1.340527e-30	NaN	NaN
C(property_type)	NaN	5.297193e-78	NaN
C(room_type)	NaN	NaN	1.834669e-24
Residual	NaN	NaN	NaN

3.2.1.2 Algorithm screening

- **Forward Selection**

Forward selection starts with no variables and adds one variable at a time. At each step, the variable that improves the model the most (based on criteria like AIC, BIC, or R-squared) is added.Stops when no significant improvement is made by adding additional variables. Finally, [‘accommodates’, ‘bathrooms’, ‘bedrooms’, ‘beds’, ‘neighbourhood_cleansed’, ‘review_scores_value’, ‘availability_365’, ‘longitude’, ‘property_type’, ‘host_is_superhost’] are screened by forward selection.

- **Backward Selection**

Backward selection starts with all candidate variables included in the model and removes one variable at a time. At each step, the variable whose removal improves the model the most (or degrades it the least) is removed. Stops when removing any further variables significantly worsens the model.-Finally, [‘accommodates’, ‘bathrooms’, ‘bedrooms’, ‘beds’, ‘review_scores_value’, ‘longitude’, ‘latitude’, ‘maximum nights’, ‘neighbourhood_cleansed’, ‘property_type’, ‘host_is_superhost’, ‘room_-type’] are screened by backward selection.

- **Stepwise Selection**

Stepwise selection is a combination of forward and backward selection methods.Then, [‘accommodates’, ‘bathrooms’, ‘bedrooms’, ‘beds’, ‘review_scores_value’, ‘neighbourhood_cleansed’, ‘property_type’, ‘room_type’] are screened by this method.

3.2.2 Model fitting

- **Linear Model**

Select the linear model as the main model for analysis. It is worth noting that we performed log on the price to improve the R square.

Table3. Comparison of different variables for linear model

Method	R-squared	Adj. R-squared	F-statistic	Log-Likelihood	AIC	BIC
Forward Selection	0.712	0.709	215.6	-4848.1	9912.0	10680.0
Backward Selection	0.712	0.709	209.7	-4846.6	9915.0	10710.0
Stepwise Selection	0.703	0.700	206.6	-4990.4	10200.0	10970.0
Manual Selection	0.721	0.718	267.4	-6567.1	13350.0	14150.0

- Forward Selection and Backward Selection provide similar R-squared values, indicating similar performance.
- Stepwise Selection has a slightly lower R-squared but is more balanced in feature selection.
- Manual Selection shows the best R-squared and is selected as optimal one.

Choose the optimal one and use SAS for further analysis:

The GLM analysis provides a robust model for predicting Airbnb prices with a high R-squared value of 0.723202, indicating that approximately 72% of the variance in prices is explained by the model. Significant predictors include the number of accommodates, bathrooms, bedrooms, and other relevant variables, all showing strong statistical significance. The detailed Type I and Type III sum of squares analysis confirms the individual importance of each predictor in the model.

GLM 过程

因变量: price

源	自由度	平方和	均方	F 值	Pr > F
模型	115	6956.234106	60.488992	320.23	<.0001
误差	14095	2662.423829	0.188891		
校正合计	14210	9618.657935			

R 方	变异系数	均方根误差	price 均值
0.723202	7.853662	0.434616	5.533933

源	自由度	I 型 SS	均方	F 值	Pr > F
accommodates	1	4135.407782	4135.407782	21893.0	<.0001
bathrooms	1	155.469353	155.469353	823.06	<.0001
bedrooms	1	63.645141	63.645141	336.94	<.0001
beds	1	4.165097	4.165097	22.05	<.0001
calculated_host_list	1	85.112999	85.112999	450.59	<.0001
availability_30	1	4.167332	4.167332	22.06	<.0001
reviews_per_month	1	27.753949	27.753949	146.93	<.0001
neighbourhood_cleans	37	1667.672587	45.072232	238.61	<.0001
property_type	68	801.296059	11.783766	62.38	<.0001
room_type	2	2.732160	1.366080	7.23	0.0007
host_is_superhost	1	8.811648	8.811648	46.65	<.0001

源	自由度	III 型 SS	均方	F 值	Pr > F
accommodates	1	65.8756159	65.8756159	348.75	<.0001
bathrooms	1	108.5130346	108.5130346	574.47	<.0001
bedrooms	1	56.1623142	56.1623142	297.33	<.0001
beds	1	1.0104785	1.0104785	5.35	0.0207
calculated_host_list	1	3.4524023	3.4524023	18.28	<.0001
availability_30	1	113.4749359	113.4749359	600.74	<.0001
reviews_per_month	1	26.3152208	26.3152208	139.31	<.0001
neighbourhood_cleans	37	949.6550188	25.6663519	135.88	<.0001
property_type	67	223.5019147	3.3358495	17.66	<.0001
room_type	2	2.2320693	1.1160347	5.91	0.0027
host_is_superhost	1	8.8116478	8.8116478	46.65	<.0001

Figure 24: GLM Process and Analysis for Airbnb Price Prediction

3.2.3 Model comparison

Table4. Model Performance Comparison

	Model	Mean Squared Error	R-squared
1	Linear Model	238669.848728	0.38953
2	Random Forest	8952.354311	0.60965
3	Linear Model (Log)	0.188891	0.723202
4	Random Forest (Log)	0.197569	0.714276
5	XGBoost (Log)	0.18433	0.733423

Note: The linear model, xgboost and random forest after log transformation are all fitted using the best filtered variable combination.

- **Interpretability and Simplicity:** Linear Model with Log Transformation maintains high interpretability and simplicity while significantly improving performance over the non-transformed version. It's suitable for scenarios where understanding the model and explaining it to stakeholders is crucial.
- **Balance Between Performance and Interpretability:** Random Forest with Log Transformation offers a good balance between enhanced performance and moderate interpretability. Feature importance can still provide insights into which features are most influential.
- **Predictive Performance:** XGBoost with Log Transformation excels in predictive accuracy but at the expense of interpretability and simplicity. This model is ideal when the highest accuracy is required, and resources are available to manage its complexity.

3.3 Model diagnosis

3.3.1 Residual plot

- Residuals vs. Predicted Values: Although the predicted prices are concentrated, the mean of the residuals is close to zero, and the variance is roughly constant.
- QQ-plot: The deviation of the sample quantile is larger where the theoretical quantile is larger, which means that the tail of the sample quantile is thicker.

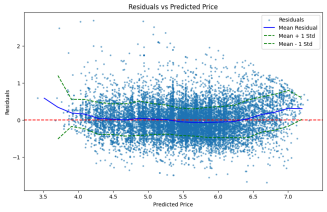

Figure 26: residual plot

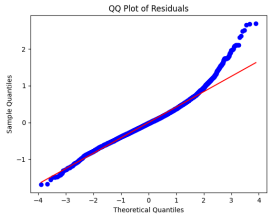

Figure 27: QQ-plot

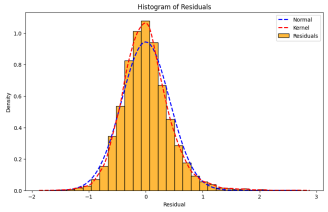

Figure 28: residual histogram

- Histogram of Residuals: Consistent with the results of QQ-plot, the residuals approximately follow a normal distribution. There is a slight deviation on the right side of the peak (high-price area).

3.3.2 Hypothesis test

The hypothesis test for residual includes tests for heteroscedasticity, normality, and independence to ensure the validity of the regression model's assumptions.

Heteroscedasticity Test: Conducted to check if the variance of residuals is constant across different levels of the predicted values.

Normality Tests: Shapiro-Wilk Test Statistic: $0.9822 \approx 1$, Kolmogorov-Smirnov Test Statistic: 0.0361. Given the sample size is 11,259 (which is greater than 5,000), the Kolmogorov-Smirnov test is more suitable for assessing the normality of the residuals here. Both tests suggest that the residuals approximately follow a normal distribution, although the distribution tails are thicker than those of a normal distribution.

Independence Test: Durbin-Watson Statistic: 1.9641. This statistic is approximately equal to 2, indicating that the residuals are essentially independent of each other.

Overall, the residual analysis supports the assumptions of the regression model, though there is a slight deviation in the normality of the residuals as indicated by the thicker tails in the distribution.

3.3.3 Multicollinearity

To address multicollinearity, we used the Variance Inflation Factor (VIF) to detect and iteratively remove variables with the highest VIF until all remaining variables, except the constant term, had a VIF less than 10.

- Before Removal: R-squared: 0.721 Adjusted R-squared: 0.719 Test R-squared: 0.7207
- After Removal: R-squared: 0.718 Adjusted R-squared: 0.716 Test R-squared: 0.7192

The removed variables include:

- `property_type_Shared room in home`,
- `property_type_Entire rental unit`,
- `room_type_Private room`,
- `neighbourhood_cleansed_Sydney`

It was observed that these variables were highly correlated with the retained variables.

For instance: `property_type_Shared room in home` was correlated with `room_type_shared_room`. `property_type_Entire rental unit` was correlated with `room_type_Entire_home/apt`. `room_type_Private room` had strong correlations with various subcategories of private rooms under `property_type`. By removing these variables, we effectively reduced multicollinearity without significantly impacting the model's explanatory power, as indicated by the slight changes in R-squared values.

4 Conclusion

Our analysis of Airbnb listings in Sydney offers several key insights:

1. **Neighborhood Analysis**: Manly has the highest number of listings, and most listings are concentrated in the bay area. The priciest accommodations are also predominantly located in this region. Safety varies by neighborhood, with theft being the most common crime, and safer neighborhoods generally have higher review ratings.

2. **Accommodation Types**: Entire homes/apartments dominate the market, while hotels are the least economical choice on Airbnb. Property type analysis reveals a mix of accommodations, with some misclassification observed.

3. **Pricing Patterns**: Average prices fluctuate throughout the year, with noticeable declines during certain periods, likely due to unupdated future prices.

4. **Clustering Analysis**: Four distinct clusters of properties were identified based on per capita metrics of bedrooms, bathrooms, and beds. Each cluster reflects different price levels and property characteristics, offering insights into the diversity of the market.

5. **Predictive Modeling**: A linear regression model was developed to predict listing prices. Through careful variable selection and transformation, the model achieved a high R-squared value, indicating strong predictive performance. Model diagnostics confirmed the validity of the assumptions, although some deviation in residual normality was observed.

Overall, this study provides valuable information for both Airbnb hosts and potential guests. Hosts can optimize their pricing strategies based on property features and neighborhood characteristics, while guests can make informed decisions by considering safety, pricing, and accommodation types. The clustering and predictive modeling approaches used here can be applied to other regions to gain similar insights into the Airbnb market.

[1] S. Lloyd, "Least squares quantization in PCM," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[2] P. Rousseeuw, "Rousseeuw, P.J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Comput. Appl. Math. 20, 53-65," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.

[3] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Journal of the royal statistical society. series c (applied statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[4] D. Marutho, S. Hendra Handaka, E. Wijaya, and Muljono, "The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News," in *2018 International Seminar on Application for Technology of Information and Communication*, 2018, pp. 533–538. doi: 10.1109/ISEMANTIC.2018.8549751.