CO3093 - Big Data and Predictive Analytics

CW Assignment Report

Mohammed Ahmed Yousif Abuharira
maya2

March 26, 2025

Table of Contents

1 Introduction

This report details the analysis of the London Airbnb listings dataset, which includes cleaning the dataset, exploring the dataset, linear regression modelling, K-Means clustering, cluster-based local regressors, and an improved model to predict Airbnb prices per listing.

2 Cleaning the Data

2.1 Shape of the dataset

The initial shape of the dataset was (66679, 31) which included missing values, duplicates, and outliers.

2.2 Transforming Variables

After inspecting the dataset, the following columns were transformed:

- **price**: price: Removed the dollar sign $ and converted to numerical.
- **host_response_rate, host_acceptance_rate**: Removed percentage sign % and converted them to percentages (numerical features).
- **host_is_superhost**: which contains Boolean values was transformed to numerical by encoding true and false to 1 and 0 respectively.

2.3 Dropping Columns:

The columns below were dropped as they are unnecessary, or useless, for predicting the price of Airbnb listings:

1. **Identification columns** (`id`, `host_id`, `host_name`): These columns identify a listing (`id`) and the host (`host_id`, `host_name`), but they are not useful in predicting the price, so they were removed.
2. **Textual columns** (`name`, `description`, `bathrooms_text`): These columns contain information that is not useful for the task at hand. For instance, `bathrooms_text` lists the number of bathrooms, which is already in the column `bathrooms`.
3. **Dates** (`calendar_last_scraped`, `first_review`, `last_review`): These are not useful for predictions. For instance, `calendar_last_scraped` is the same for all listings ($\pm$ 8 days).
4. `Amenities column`: Although it might seem important at first, this column was dropped due to the following reasons:
    - Inconsistent naming: the same amenity is named differently across listings.
    - Large size of unique elements (approximately 9000), which increases dimensionality.
    - Elements as sentences, sometimes with multiple types of amenities, i.e., "Children 2019s books and toys for ages 0-2 years old, 2-5 years old", complicating data pre-processing.
5. **Location variables** (`latitude`, `longitude`): Dropped as the column `neighbourhood` already contain valuable information about the location of Airbnb listings.

The removal of these columns changed the shape of the data set to 66679 rows and 18 columns.

2.4 Duplicate Removal

The dataset contained 33 duplicates, however, after dropping the columns in **Section 2.3**, the number of duplicates increased to 442, which is expected given that those columns could have been distinguishing rows from each other. All of the 442 duplicate rows were removed, changing the shape of the dataset to 66237 rows and 18 columns.

2.5 Handling Missing Values

Before handling missing values, empty strings/lists in categorical variables were removed and replaced them with Nan. Then, missing values, rows with NaN, were removed from the dataset since they (23517 in the dataset) cause issues, as many algorithms do not handle them well. Dropping these values reduced the shape of the data frame to 42720 rows and 18 columns. **Table 1** below displays the missing values per column, for the remaining columns. For a table of missing values per column in the original dataset, refer to **Appendix A**.

Table 1: Missing Values in the Dataset

Column	Missing Values	Percentage
host_response_rate	10,175	15.26
host_acceptance_rate	7,146	10.72
host_is_superhost	380	0.57
host_listings_count	1	0.001
host_total_listings_count	1	0.001
bathrooms	5,329	7.99
bedrooms	101	0.15
beds	5,400	8.10
price	5,261	7.89
review_scores_rating	14,736	22.10

2.6 Outlier Detection and Removal

Two methods were tested to detect and handle outliers in all numerical columns: IQR and Z-score. The chosen method was the IQR method given that the data is skewed and thus the Z-score method, which assumes normal distribution, will fail to detect outliers. **Figure 1** below displays the distribution of numerical columns, and it can be seen that almost all of them skewed (note the price plotted is on the log scale, which is covered in **Section 2.7**).

2.6.1 IQR Method:

Which is a method to detect outliers that fall outside the lower and upper bounds below:

$$\text{Lower bound} = Q_1 - 1.5 \times IQR$$

$$\text{Upper bound} = Q_3 + 1.5 \times IQR$$

Where Q_1 represents the first quartile, Q_3 represents the third quartile, and IQR is the interquartile range ($IQR = Q_3 - Q_1$).

This method can be used to remove or cap outliers, the former use was applied in this section. Using this method reduces the data aggressively from 42720 rows, after removing missing values, to 18165 rows (57.5% reduction). This is to be expected as the numerical variables are highly skewed, as stated earlier, and thus removing outliers will result in such a reduction. This reduction in data is addressed in the *improved model* part of the report in **Section 5.3**.

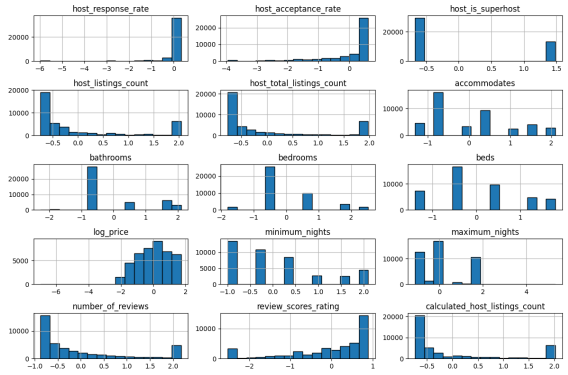

Figure 1: Distributions of Numerical Variables

2.7 Addressing Skewed Numerical Variables

The log of the `price` column was taken, to handle skewness, and it was renamed to `log_price`. The skewness in the rest of the numerical variables is addressed in **Section 5.3.3**.

2.8 Data Normalisation

Following outlier removal, the numerical columns in the dataset were normalised using Standard Scaler, which scales the data to have a mean of 0 and a standard deviation of 1. Given that outliers were removed, i.e., they have no effect on the mean or the standard deviation, the Standard Scaler is a suitable as it will not change the shape of the data and focus on centring and scaling the data.

3 Data Exploration

This section details the different explorations on the data, which allow us to understand the data in great detail.

3.1 Prices Across Neighbourhood

The bar plot in **Figure 2** was created, which shows the average Airbnb prices across different neighbourhoods. In **Figure 2**, it can be seen that average prices vary significantly across different neighbourhoods, with places such as Westminster, the City of London, and Kensington and Chelsea having the highest average prices, which is consistent with the general knowledge of London. This indicates that the variable neighbourhood have a strong influence on the price of a listing.

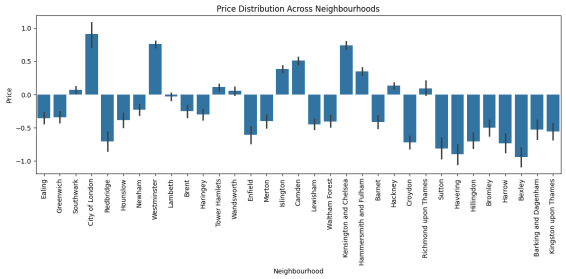

Figure 2: Average Price Across Neighbourhoods

3.2 Visualize the prices across number of possible tenants

Now we want to explore if the price increases with the number of possible tenants. **Figure 3** displays a scatter plot between the variables **log_price** and **accommodates** this was done by grouping properties by the number of possible tenants and calculating the mean price for each group. It can be seen that there's a positive correlation between the price of a listing and the number of people a property can accommodate. This indicates that the variable accommodates have a high correlation with the variable price and could be a strong predictor of it.

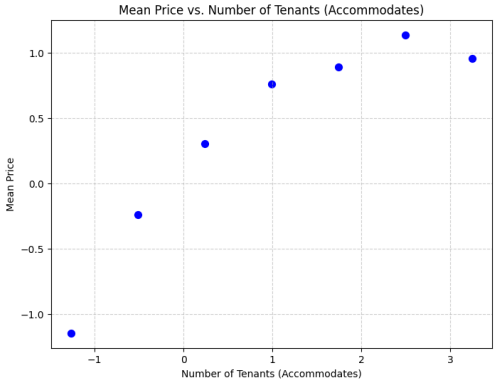

Figure 3: Mean Price vs. Number of Tenants (Accommodates)

3.3 Does the average review rating affect the process?

This question could be answered by grouping the data points by the review score rating and calculating the mean price for each group. In **Figure 4**, it can be seen that there is a weak correlation between the variables. This might indicate that the variable does not directly affect the price, or there is a complex relationship between the variables that is not visible on the plot.

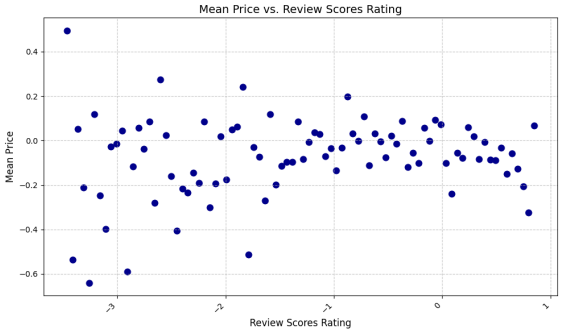

Figure 4: Mean Price vs. Review Scores Rating

3.4 Distribution of mean price by room type

Figure 5, which displays the distribution of room types in the dataset, shows that the categories: `entire home`, and `private rooms` are the most common room type in the dataset with `shared room` and `hotel room` being the least common, which is expected for Airbnb listings.

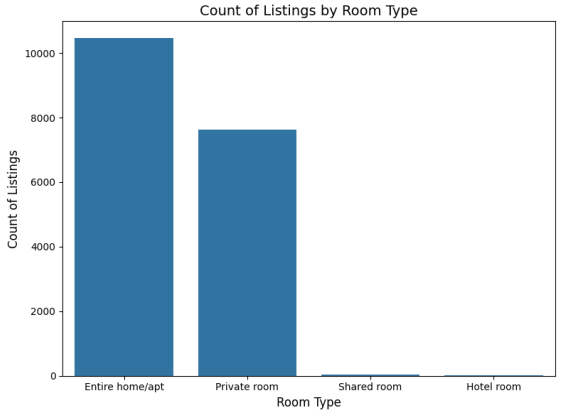

Figure 5: Distribution of Room Type in Listings

3.5 Distribution of mean price by property type

Figure 6 displays the top ten property types with the highest mean price, which confirms that larger properties, e.g., `entire condo`, `home`, `rental unit`, etc. have higher prices.

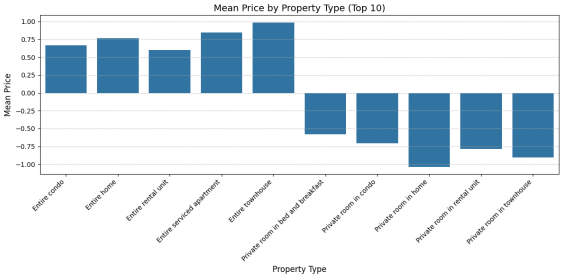

Figure 6: Mean Price by Property Type (Top 10)

3.6 Correlation Matrix

The correlation matrix of the remaining numerical features, after cleaning, was created and can be seen in **Figure 7** and the top five features with the highest correlation with the price were identified as the following in **Table 2**:

Table 2: Top correlations with price

Feature	Correlation with price
accommodates	0.639065
bedrooms	0.436436
beds	0.322540
minimum_nights	0.182556
bathrooms	0.134007

It can be seen that the number of occupants, bedrooms, beds, minimum nights, and bathrooms have high correlations with the pricing of a listing. This is to be expected and not out of the ordinary as those are features that drive the price of properties in the rental market in general.

Finally, the low correlation (**0.028146**) between the `review_scores_rating` and the `price` is shown in **Figure 7**. One possible cause of this low correlation is that most of the review scores in this dataset are positive, this can be seen from the distribution of `review_scores_rating` in **Figure 1**, where most of the scores are concentrated at the higher end of the rating (5 on the original scale), resulting in a weak positive correlation between the two variables. This takes us back to the question in **Section 3.3**, which could be answered now with confidence, no, the review score rating does not affect the pricing process, at least not in this dataset.

3.7 Scatter Matrix plot

The scatter matrix, **Figure 8**, reveals positive relationships between `log_price` and `accommodates`, `bedrooms`, `beds`, and `bathrooms`, consistent with the correlation matrix, suggesting that larger properties tend to be more expensive.

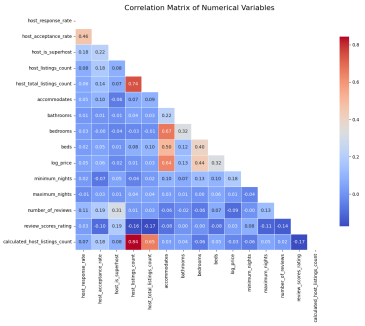

Figure 7: Correlation matrix plot of Key Variables

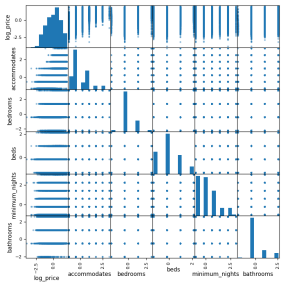

Figure 8: Scatter matrix plot of Key Variables

4 Linear Model

4.1 Model Building

A basic predictive model to predict the prices of Airbnb listings was created using linear regression. The model uses both numerical and categorical predictors:

1. **Numerical predictors:** These are the top five features with the highest correlations with the price.

2. **Categorical features:** These are features that were discovered during the exploration of the dataset, such as `property_type`, `room_type`, and `neighbourhood`. Additionally, `host_is_superhost` was selected as it is logical that superhosts can charge higher prices, which was confirmed during model training.

The dataset was split into 80% for training and 20% for testing. Categorical features were transformed into numerical features using one-hot encoding to ensure compatibility with the model.

4.2 Evaluation

To evaluate the model, the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 score were calculated on both the test set and using cross validation (5 folds). The evaluation metrics, **Table 3**, indicate a very good level of performance of the linear model.

Table 3: Linear Model Evaluation

Model Evaluation on Test Set	
Mean Absolute Error (MAE)	0.4144
Root Mean Squared Error (RMSE)	0.5302
R^2 Score	0.7178
Evaluating Model using Cross-Validation (5-fold)	
Mean R^2 Score	0.7116
Standard Deviation R^2 Score	0.0037
Mean RMSE	0.5369
Mean MAE	0.4200

4.3 R^2 Score

The R^2 score of 0.71 on both the test set and in cross-validation suggests that the model is very good at explaining the variance in the target variable, `price`. Furthermore, the low standard deviation of the R^2 score in cross-validation demonstrates the consistency of the model across different data folds.

4.4 RMSE and MAE Scores

The RMSE of 0.53 on the logarithmic scale indicates a prediction error of around 0.53 price units on the logarithmic scale. Likewise, the MAE of 0.42 suggests an average absolute error of 0.42 price units on the logarithmic scale. These values represent a manageable level of prediction error, especially for a linear model on such complex real-world data.

4.5 Residuals Analysis

Finally, the following histogram of residuals and Q-Q plot, **Figures 9 and 10**, show that the residuals roughly follow a bell shape centred around 0, suggesting that the model errors are normally distributed with no apparent pattern.

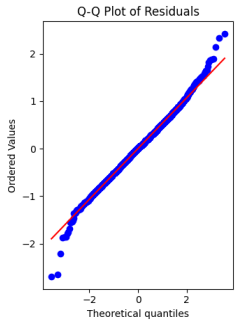

Figure 9: Q-Q Plot of Residuals (Log Scale)

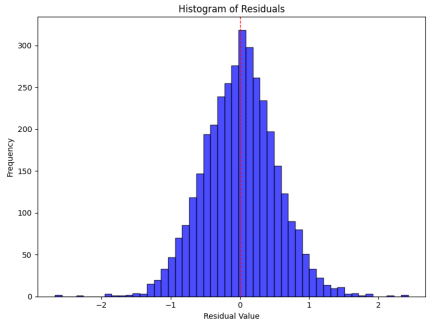

Figure 10: Histogram of Residuals (Log Scale)

5 Part 2: Improved model

5.1 Clusters

The data was clustered using numerical features that indicate size and capacity these are: (accommodates, bedrooms, beds, bathrooms) and the host's total listings count, as indicate the experience of the host, and the minimum nights required as it differentiates long and short term listings. K-Means algorithm was used to group the data into 4 groups which can be seen in **Figure 11**. The optimal number of clusters was determined with the help of the elbow method, which is a plot of different cluster sizes and their cost (or inertia). As it could be seen in **Figure 12** the elbow is between 4 and 8. Starting from 4, different values were tested and the number of cluster chosen was 4 as it created more meaningful clusters compared to other low-cost cluster sizes.

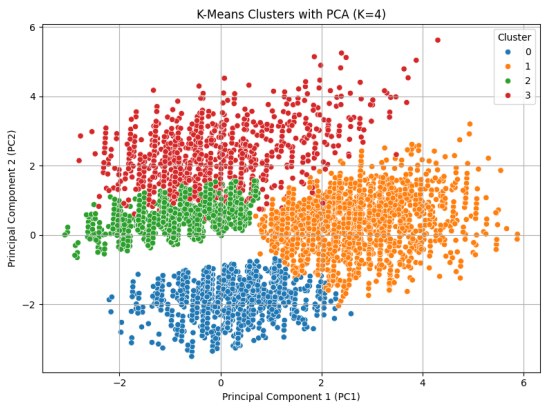

Figure 11: K-Means Clusters Visualized with PCA

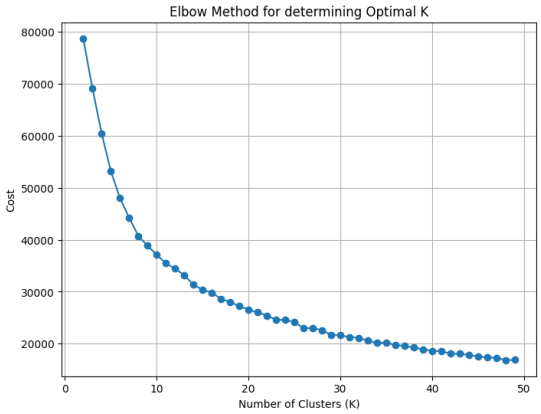

Figure 12: Elbow Method for determining Optimal K

5.1.1 Distributions of Features in Clusters

This section covers the characteristics of the clusters using **Table 4**, which displays the distribution of numerical features in each cluster. Additionally, the distribution of categorical features in each cluster will also be covered in this section. For detailed information on the distributions of the variables neighbourhood, property type, and room type (categorical features) in each cluster refer to the tables in **Appendix B**.

Table 4: Standardized Numerical Features Means Per Cluster

Cluster	accommodates	bedrooms	beds	minimum_nights	bathrooms	host_total_listings_count
0	-0.383846	-0.410315	-0.296610	1.169807	-0.190029	-0.235185
1	1.374076	1.567488	1.033248	0.256010	0.733322	-0.131567
2	-0.437691	-0.428624	-0.343536	-0.645556	-0.202292	-0.275922
3	0.177722	-0.250308	0.231958	-0.040850	-0.112382	2.394005

Cluster 0 (size = 4075): The second largest cluster. This cluster groups properties with fewer than average beds, bedrooms, and bathrooms, i.e., small properties. Listings in this cluster have a longer minimum stay, the highest among all clusters, indicating long-term rentals. Moreover, properties in this cluster are managed by hosts with fewer than average total listings, suggesting that the hosts could be individual owners rather than professional companies. Another distinction is the room type in this cluster as it is made up of **39%** private room and **61%** Entire home/apt. Finally, the price of listings in this cluster is slightly below average (**Table 5**), which is consistent with the analysis that this cluster groups small properties.

Cluster 1 (size = 3719): The second smallest cluster. This cluster groups properties with highly above-average capacity, bedrooms, and beds, with an above-average number of bathrooms and minimum stay but below-average host total listings. They represent large accommodations suitable for renters with a high budget, extremely above average price (**Table 5**), seeking maximum space, best for events such as family gatherings, weddings, etc.

Cluster 2 (size = 8757): The largest cluster. This cluster groups smaller properties with below-average means in all features. Property types in this cluster are diverse including entire

rental units (**50%**), entire condo (**23.18%**), entire home (**15.68%**), and some other types with lower percentages. Properties in this cluster represent budget-friendly listings with short stays, likely tailored towards renters on a budget or solo travellers, confirmed by the lower-than-average price in **Table 5**.

Cluster 3 (size = 1614): The smallest cluster. This cluster groups properties with slightly above-average capacity and beds, slightly below-average bedrooms, minimum nights, and bathrooms, and well above-average host total listings. Property types in this cluster are mostly rental units (**45.7%**), houses (**20.6%**), and apartments (**19.5%**). Properties in this cluster represent medium-sized properties managed by professionals, explained by the high listing counts, who are efficient in managing space in medium-sized properties, accommodating more people despite below-average bedrooms in said listings. Finally, the slightly above-average price in **Table 5** is to be expected for medium-sized properties in London.

Table 5: Price Summary (Log-Normalized)

Cluster	log_price
0	-0.0997271
1	0.930042
2	-0.3687
3	0.109211

5.2 Local Regressors Based on Clusters

The dataset was sub setted by cluster and using the same numerical and categorical features, the linear regression model in **Section 4** was trained on each subset.

5.2.1 Evaluation

Table 6 displays the evaluation metrics for all regressors discussed below:

Cluster 0: Strongest local model with a mean R^2 (CV) of 0.6824 and mean RMSE of 0.5331, and a mean MAE of 0.4166, with a performance close to the basic model's performance, with slightly better errors. This relatively strong performance is likely due to the consistent pricing within small properties.

Cluster 1: The weakest local regressor with a much lower cross-validation (CV) mean R^2 of 0.3778, and slightly better mean MAE (CV) of 0.3862 and mean RMSE of 0.4934 (CV) compared to the basic model. This poor performance may be due to the complex pricing in expensive listings, where there are other factors affecting the pricing, resulting in complex relationships that the linear model is failing to learn. Another reason for the poor performance might be the absence of features that drive the prices of premium listings such as exclusivity, luxury amenities, etc.

Cluster 2: The performance of the local regressor in this cluster is lower than the basic model, but still explains a good percentage of the variance in the price (mean R^2 (CV) of **0.6298**). Additionally, the errors are slightly worse with a 0.5448 mean RMSE (CV) compared to 0.5369 in the basic model, and 0.4242 mean MAE (CV) compared to 0.42 in the basic model. This may be because budget-friendly listings are diverse, which creates variability in prices, and the linear model is not learning this variability in the same way as the basic model.

Cluster 3: The second strongest local regressor with a mean R^2 (CV) of 0.6614 and mean RMSE of 0.5693, and a mean MAE of 0.4437, with a performance not too far off the basic model performance, with slightly worse errors. The relatively strong performance suggests that professional hosts tend to be more consistent with pricing, with aspects such as location adding some variability to their pricing and thus lowering the performance of the regressor compared to the global model.

Note: The difference between the R^2 from the test set and the cross-validation for the regressors in clusters **0** and **3**, **Table 6**, suggests slight overfitting in these regressors, especially the one on cluster **3**, however, the difference is not significant, only 0.04 in the worst case.

5.2.2 Conclusion and Recommendations

Although clustering did not lead to improvements over the basic model, it revealed valuable insights that could drive the development of a much stronger model. It is recommended to develop a non-linear model to address the poor performance in Cluster 1, as it was concluded that the poor

Table 6: Evaluation Metrics for Local Linear Regression Models per Cluster

Cluster	RMSE (Test Set)	R^2 Score (Test Set)	Mean R^2 (CV)	Mean RMSE (CV)	Mean MAE (CV)
0	0.5057	0.7139	0.6824	0.5331	0.4166
1	0.5118	0.3713	0.3778	0.4934	0.3862
2	0.5475	0.6217	0.6298	0.5448	0.4242
3	0.5507	0.7052	0.6614	0.5693	0.4437

performance in that cluster could be due to non-linear pricing relationships, and such model could address these issues leading to better performance.

5.3 Improved Model

A Random Forest Regression model was chosen as the improved model. Before training the improved model, some aspects of data preprocessing were changed. This section covers changes in the cleaning process, the selection of predictors, and finally the evaluation of the improved model and recommendations for future improvements.

5.3.1 Handling Missing Values

First, imputation was tested in handling missing values, which resulted in more data being kept, however, the increase in data did not lead to improvements in the model's performance, in fact, it led to longer training time, as such, missing values were dropped, same as in **Section 2.5**.

5.3.2 Handling Outliers

Handling outliers for the improved model was done differently, instead of using the IQR method, described in **Section 2.6.1**, to remove them, it was used to limit, cap, them to the values of the lower and upper bounds, which resulted in 42720 rows kept (i.e., the number of rows left after removing missing values). Capping outliers resulted in some duplicates, 19 rows, and those were also removed.

5.3.3 Handling Skewed Values

Here the skewed variables that were not handled earlier in the report, in **Section 2.7**, are addressed. Skewed numerical values were power-transformed using the `Yeo-Johnson` method, as it could handle left and right skewed distributions. Power-transforming the skewed variables increased the correlations with the **price** variable, see **Table 7** or **Appendix C** for an image of the correlation matrix. This resulted in variables such as **host_total_listings_count** moving up the correlation table, and the variable **minimum_nights** being outside the top-five features compared to **Table 2**.

5.3.4 Selecting Predictors

For the numerical predictors, the new top five features in the correlation table were selected as the numerical predictors (see **Table 7**). As for the categorical predictors, they remain the same as in **Section 4**.

Table 7: Top correlations with price

Feature	Correlation with log_price
accommodates	0.712364
bedrooms	0.511140
beds	0.457450
bathrooms	0.303817
host_total_listings_count	0.237461

5.3.5 Evaluation

Table 8 below shows the results of the improved model, which show a significant improvement (over 10% in percentage difference) in the mean R^2 (CV) score over the basic linear regression model. Moreover, the improved model resulted in a 13% reduction in the mean RMSE (CV) and a 19% reduction in the mean MAE (CV), meaning less errors and improved prediction accuracy.

Similar to figures in **Section 4.5**, the histogram of residuals and Q-Q plot, **Figures 13 and 14**, show that the residuals roughly follow a bell shape centred around 0, with slightly better normality compared to the model in **Section 4**.

Finally, the Random Forest model significantly outperforms the basic linear model, doing a much better job at capturing the non-linear relationships with the price, which confirm that the conclusion in **Section 5.2.2** was accurate and the Random Forest model is better suited for the Airbnb dataset.

Table 8: Improved Model Evaluation Metrics

Model Evaluation on Test Set	
Mean Absolute Error (MAE)	0.3343
Root Mean Squared Error (RMSE)	0.4506
R^2 Score	0.7999
Evaluating model using Cross-Validation (5-fold)	
Mean R^2 Score	0.7901
Standard Deviation R^2 Score	0.0054
Mean RMSE	0.4580
Mean MAE	0.3383

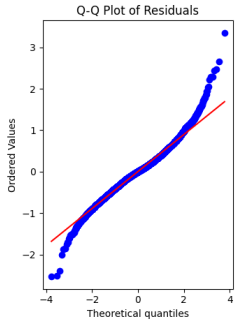

Figure 13: Q-Q Plot of Residuals (Log Scale)

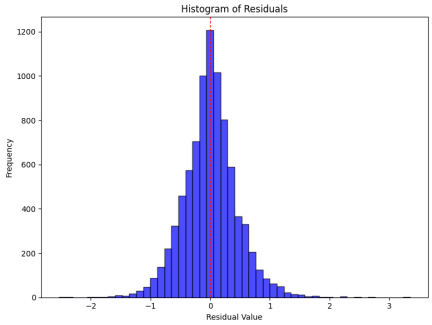

Figure 14: Histogram of Residuals (Log Scale)

5.3.6 Recommendations

Although earlier recommendations to train a non-linear model to improve the basic model worked, additional collection of data is needed for further improvements such as the listings' proximity to tourist attractions, public transport, restaurants, parks, etc., as this information may result in better modelling of the relationships with the price of a listing and lead to higher accuracy.

6 Conclusion

The London-Listings dataset was cleaned, explored, clustered, and predictive models were created. Clustering revealed that there may be non-linear pricing relationships in the premium properties Airbnb market. This was addressed in the improved model section by using a random forest model that improved significantly over the basic mode. As to how this improved model may be used, it is recommended to create a dashboard with the improved model as the pricing engine. This will help hosts increase their revenue by changing the values of some variables and getting price estimates from the model enabling them to answer questions like "How much would adding two beds affect my annual revenue?". Finally, the model could be integrated into the listing creation process, by Airbnb, to provide price recommendations to hosts listing new properties.

Appendix

A Missing Values Per Column In the original Dataset

#	Column	Missing Values	Percentage (%)
1	description	1,956	2.93
2	host_name	1	0.0015
3	host_since	1	0.0015
4	host_response_rate	10,175	15.26
5	host_acceptance_rate	7,146	10.72
6	host_is_superhost	380	0.57
7	host_listings_count	1	0.0015
8	host_total_listings_count	1	0.0015
9	bathrooms	5,329	7.99
10	bathrooms_text	100	0.15
11	bedrooms	101	0.15
12	beds	5,400	8.10
13	amenities	136	0.20
14	price	5,261	7.89
15	first_review	14,736	22.10
16	last_review	14,736	22.10
17	review_scores_rating	14,736	22.10

Table 9: Summary of Missing Values in the Dataset

B Distribution of Categorical Features per Cluster

Table 10: Distribution of neighbourhoods across London boroughs by cluster (%)

Borough	**Cluster 0**	**Cluster 1**	**Cluster 2**	**Cluster 3**
Barking and Dagenham	0.69	1.32	1.03	0.87
Barnet	1.87	2.85	2.75	2.79
Bexley	0.49	0.83	1.20	0.56
Brent	2.99	2.77	3.23	3.84
Bromley	0.98	1.40	1.54	0.99
Camden	5.42	5.73	4.26	7.06
City of London	0.54	0.16	0.21	0.56
Croydon	1.82	2.12	2.71	0.68
Ealing	2.72	2.77	2.81	1.80
Enfield	0.96	1.34	1.53	0.68
Greenwich	2.50	2.58	2.93	2.79
Hackney	9.96	7.15	6.39	6.01
Hammersmith and Fulham	4.56	4.95	3.73	4.71
Haringey	2.87	2.80	3.21	0.99
Harrow	0.96	0.65	0.95	0.62
Havering	0.42	0.38	0.77	0.37
Hillingdon	0.69	0.81	1.53	1.30
Hounslow	1.37	1.61	2.03	1.73
Islington	6.40	5.43	4.56	7.00
Kensington and Chelsea	5.30	4.76	3.65	6.57
Kingston upon Thames	0.93	0.97	1.19	0.99
Lambeth	6.82	6.86	6.44	4.21
Lewisham	3.85	3.01	3.70	2.11
Merton	1.79	1.88	2.57	1.30
Newham	2.36	3.50	3.76	2.91
Redbridge	0.66	0.75	1.24	0.99
Richmond upon Thames	1.28	1.96	1.87	1.30
Southwark	6.09	5.97	6.31	4.28
Sutton	0.56	0.56	0.59	0.68
Tower Hamlets	6.63	6.61	6.87	8.86
Waltham Forest	2.23	3.01	2.91	2.23
Wandsworth	5.55	5.94	5.61	4.21
Westminster	7.73	6.56	5.90	14.00

Table 11: Distribution of property types by cluster (%)

Property Type	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Entire rental unit	37.42	50.07	24.87	45.72
Private room in home	14.26	2.45	22.69	15.37
Private room in rental unit	12.54	1.13	20.04	10.22
Entire condo	17.47	23.18	9.56	12.64
Private room in condo	8.05	0.91	11.41	4.58
Entire home	1.96	15.68	1.75	1.80
Private room in townhouse	2.28	0.30	2.31	1.18
Entire loft	0.91	0.70	0.27	0.31
Entire townhouse	0.39	2.12	0.09	0.25
Entire serviced apartment	0.61	1.26	0.71	1.43
Private room in loft	0.34	0.11	0.19	0.06
Private room in guest suite	0.47	0.03	0.42	0.31
Other types	3.30	2.06	5.69	6.13

Table 12: Distribution of room types by cluster (%)

Room Type	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Entire home/apt	60.83	94.70	39.23	63.63
Hotel room	0.00	0.00	0.16	0.25
Private room	39.12	5.27	60.19	35.63
Shared room	0.05	0.03	0.42	0.50

C Correlation Matrix for the improved model

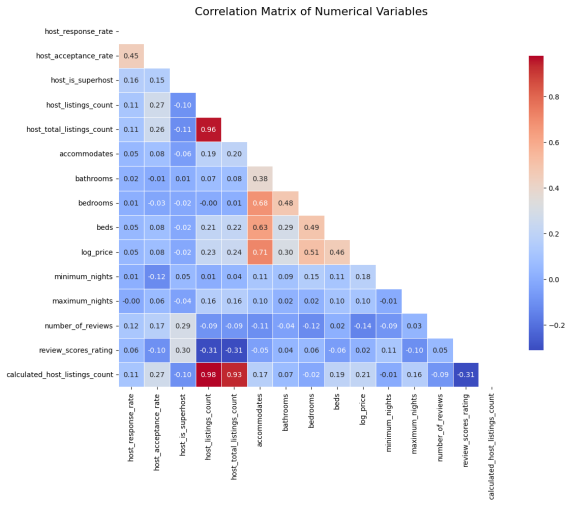

Figure 15: Correlation matrix plot of Key Variables