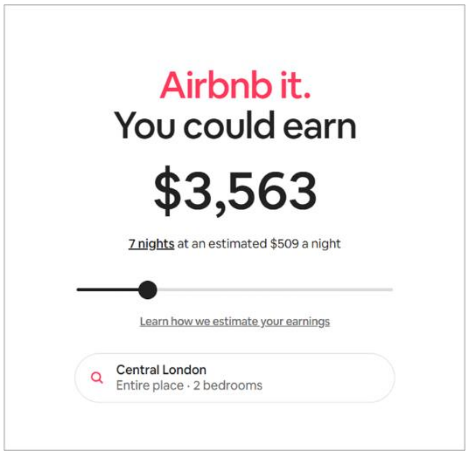
Project Report: Airbnb London Price Predictor to boost inventory of listings

Section B | Team 6 | Mahin Rahman, Denisse Castillo, Siying Ren, Yao Mu, Qingci Jiang

Executive Summary: Airbnb aims to attract potential hosts in London by providing accessible revenue insights. To achieve this, we developed a regression model to predict prices for listings, incorporating key attributes such as location, property space, amenities, and availability. Although Random Forest emerged as the most accurate model based on mean absolute error (MAE) and indicated that price was majorly influenced by distance to landmarks, amenities and neighborhood, it exhibited a concerning level of deviance, with an MAE of 47 pounds against an average listing price of 200 pounds. Consequently, we determined that the model is not suitable for deployment in its current state. We recommend flagging and separating outliers during production and exploring more robust modeling options to improve predictive accuracy.

Business Objective: Airbnb faces a challenge in attracting potential hosts in London due to a lack of accessible revenue insights. Therefore, it needs to develop an interactive price prediction tool to help potential Airbnb hosts in London estimate their property's earning potential. By offering hosts valuable information about potential earnings, Airbnb can support their decision-making process and encourage more property owners to list their homes on the platform, thereby expanding its inventory of listings.

Figure 1: Example of price prediction tool



Airbnb it.
You could earn
$3,563
7 nights at an estimated $509 a night

Learn how we estimate your earnings

Central London
Entire place · 2 bedrooms

Business Question: How can we predict the potential revenue of Airbnb listings in London to provide potential hosts with actionable insights and drive new property listings?

Action: We aim to build a regression model to predict per night Airbnb prices in London for calculating the earning potential. The model will consider various property attributes, including but not limited to: Location (e.g., proximity to landmarks, neighborhood), Property space (e.g., number of bedrooms, bathrooms), Amenities offered (e.g., WiFi, kitchen, dedicated workspace), Availability (e.g., days booked).

Dataset Overview: Each record of the dataset contains details of Airbnb listings in London including attributes of the property, host, and price per night. The details of the dataset are as follows. Further details about the data types of fields are mentioned in Appendix 1.

- Number of rows: 96,183
- Number of columns: 75
- Time: data scraped on 6th to 11th September 2024.

A map visualization of the locations of Airbnb listings (Appendix 2) also shows that majority of the listings are positioned in Central London where many key tourist attractions are present. Moreover, a summary of the dataset shows a median of 2 listings per host, which suggests that many Airbnb hosts in London are likely operating professionally, using the platform as a business. However, the majority of the variables in this initial dataset, including prices are character type, indicating the need to transform relevant fields to prepare it for our regression.

Data Source: https://insideairbnb.com/

Data Preparation (Details in Appendix 3):

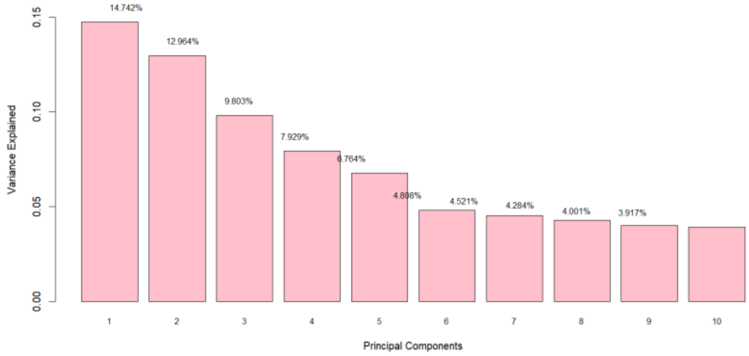
- Deleting Fields (Part 1): We removed 36 columns from the dataset, including identifiers, irrelevant URLs, and unsuitable text fields. Date fields, duplicates, sparse data, and multicollinear fields were also discarded. We kept short-term availability fields while excluding `availability_365`.

- Transforming and Creating Fields: Categorical variables were created from `neighbourhood_cleansed`, `property_type`, and `room_type`. Missing values in `bathrooms`, `bedrooms`, and `beds` were set to zero. We transformed `amenities` into count and dummy variables. Distances to key London landmarks were calculated.
- Deleting Fields (Part 2): The `property_type` field was removed due to redundancy, and fields like `latitude` and `longitude` were discarded after feature engineering.
- Accounting for Missing Values: We identified around 33,000 missing values in the `price` column and retained over 60,000 records after removing records with missing values. Other fields with significant missing values were discarded, resulting in 62,024 records and 32 columns.
- Creating Train and Test Dataset: The dataset was split into 90% training and 10% testing, confirming a representative split through similar mean `price` values and visualization variables through `ggpairs`. Box-plot visualizations of the prices in our training and test datasets show outliers, which may indicate high-end properties or input errors. Instead of removing these outliers, we retain them to reflect real-world scenarios, as they can help us evaluate how well models handle such cases, ultimately improving their predictive accuracy in production.

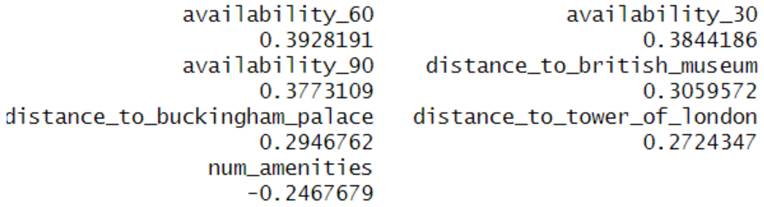
Correlation Analysis: The correlation analysis (detailed diagram in Appendix 4) reveals a moderate correlation between price and features such as the number of accommodates, beds, bathrooms, and amenities. This is logical, as these are the primary characteristics of an Airbnb listing, making them key determinants of the price. Additionally, a moderate correlation is observed between availability and distance to major landmarks, suggesting that Airbnb listings close to popular attractions may be more difficult to book.

Principal Component Analysis (PCA): The PCA results indicate that the first five principal components (PCs) explain 57% of the variance i.e. not all PCs contribute equally to explaining the variance.

Figure 2: Result of Principal Component Analysis

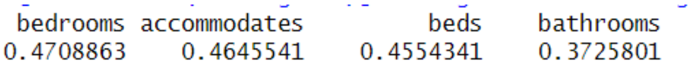


PC1 represents properties that are less luxurious and farther from key landmarks, with higher availability. This suggests they may cater to **budget-conscious travelers** seeking **affordable** options.



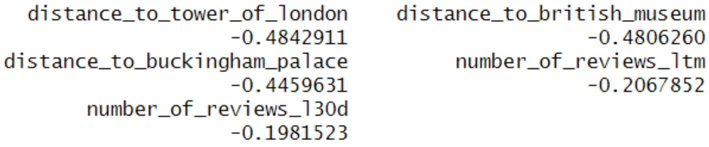
```
              availability_60            availability_30
                    0.3928191                  0.3844186
              availability_90   distance_to_british_museum
                    0.3773109                  0.3059572
distance_to_buckingham_palace   distance_to_tower_of_london
                    0.2946762                  0.2724347
                num_amenities
                   -0.2467679
```

PC2 highlights larger listings that feature more bedrooms, bathrooms, and overall accommodation capacity, making them ideal for **families or groups** looking for **spacious** options.



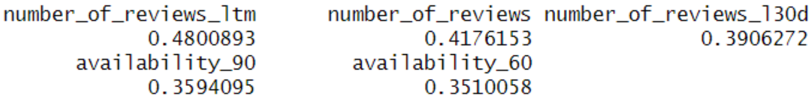
```
 bedrooms accommodates      beds   bathrooms
0.4708863    0.4645541 0.4554341   0.3725801
```

PC3 identifies **seasonal listings** near popular landmarks, which tend to receive fewer reviews. This suggests that while these listings may attract guests during peak times, they struggle to maintain consistent occupancy.



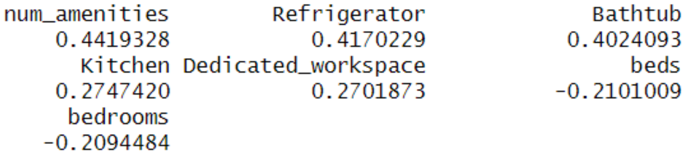
```
  distance_to_tower_of_london   distance_to_british_museum
                   -0.4842911                   -0.4806260
distance_to_buckingham_palace        number_of_reviews_ltm
                   -0.4459631                   -0.2067852
        number_of_reviews_l30d
                   -0.1981523
```

PC4 captures properties that are available year-round and receive steady reviews, likely representing **commercial** listings aimed at generating ongoing rental income.



```
number_of_reviews_ltm        number_of_reviews number_of_reviews_l30d
            0.4800893                0.4176153              0.3906272
      availability_90          availability_60
            0.3594095                0.3510058
```

Lastly, PC5 points to listings that prioritize amenities like kitchens and dedicated workspaces over sleeping arrangements, indicating a potential focus on **day-use accommodations.**



```
num_amenities        Refrigerator             Bathtub
    0.4419328           0.4170229           0.4024093
      Kitchen Dedicated_workspace                beds
    0.2747420           0.2701873          -0.2101009
     bedrooms
   -0.2094484
```

Modeling: The mathematical model for determining the earning potential of a property is as follows:

Potential Earnings= (Predicted Price per Night based on Property, Host and/ or market Characteristics) × (Number of Nights inputted in the Airbnb Predictor tool)

It’s important to note that the listing availability data in our dataset reflects the number of days a listing is available after being booked or blocked by the host. Consequently, this data does not fully capture the demand for the listing based on its characteristics. Therefore, we will refrain from predicting the occupancy rates of listings and will instead adopt an optimistic approach by assuming that all nights input into the predictor tool will be booked. To enhance accuracy, any internal Airbnb data related to occupancy rates could be incorporated.

We selected a diverse range of models to predict Airbnb prices:

1. Linear Regression
   - Pros: Linear regression is simple and interpretable, serving as a useful baseline for comparing more advanced models.
   - Cons: Its assumption of linear relationships may not capture the complexity of the Airbnb data, leading to underfitting in some cases. Plus, it can be significantly impacted by outliers.
2. CART (Classification and Regression Trees)
   - Pros: CART provides an interpretable “if-then” structure, making it easy to understand and communicate pricing strategies.

- Cons: It may be prone to overfitting, especially with deeper trees, and can be sensitive to small changes in the data.

3. Random Forest

   - Pros: This ensemble method handles complex, non-linear relationships well and is robust against overfitting, even with many features. Its ability to capture intricate relationships may be crucial for understanding pricing dynamics in a diverse market like Airbnb,

   - Cons: Random Forest can be less interpretable than simpler models, making it challenging to extract specific insights about individual predictors.

4. K-NN (K-Nearest Neighbors)

   - Pros: K-NN is intuitive and predicts prices based on the similarity of listings, making it suitable for localized pricing analysis.

   - Cons: It can struggle in high-dimensional datasets due to reliance on distance metrics, which become less meaningful with many variables.

5. Lasso Regression

   - Pros: Lasso is effective for high-dimensional data, automatically performing feature selection by shrinking irrelevant coefficients to zero, which simplifies the model.

   - Cons: It may struggle with correlated predictors, arbitrarily selecting one over the other, which could lead to loss of information.

6. Post-Lasso Regression

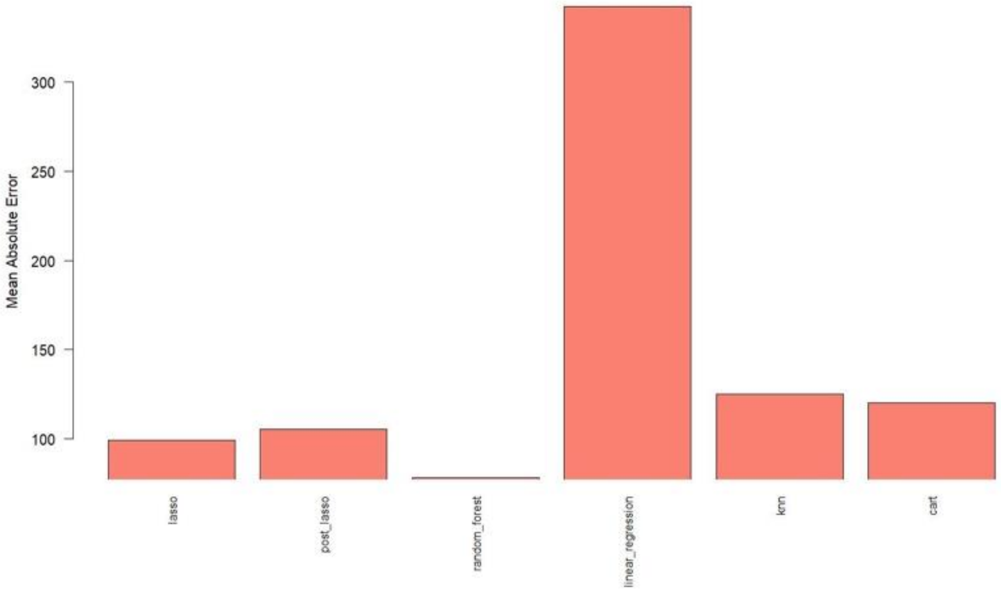
   - Pros: Post-Lasso refines the model by applying traditional regression only to the selected features, improving flexibility and potentially enhancing predictive accuracy.

- Cons: It may inherit biases from the Lasso selection process and can be sensitive to the choice of the regularization parameter.

Cross Validation: To evaluate multiple predictive model (Linear regression, CART, Random Forest, K-NN, Lasso, and Post-Lasso) and select the one that performs best in terms of predictive accuracy, we used K-fold cross validation, which helps ensure the model's generalizability to unseen data. By dividing the dataset into 10-fold sections, we ensure the model is trained and tested across multiple subsets of the data. This provides an estimate of each model's Out of Sample (OOS) performance.

Evaluation: We used **Mean Absolute Error (MAE)** to evaluate the regression models run on the training data.

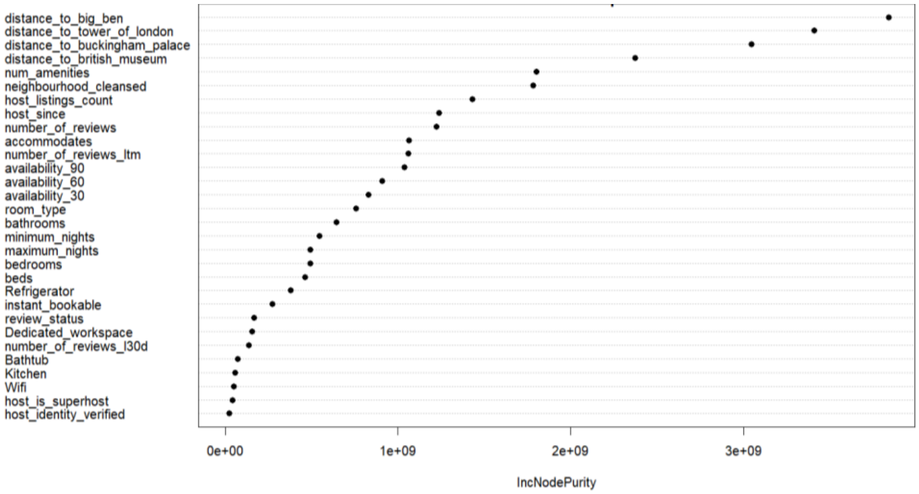
Figure 3: Mean Absolute Error (MAE) comparison



Considering MAE, we found **Random Forest is the best-performing** model with respect to the training dataset as it has the lowest MAE. It is important to highlight that the Linear Regression model has an abnormally high MAE as it is significantly impacted by outlier values and possible non-linear relationships. The significant gap between Random Forest and Linear Regression highlights the need for more sophisticated, non-linear models when predicting property prices in diverse a market like London. **We then make predictions on the test data using the Random Forest model.**

Importance Plot from Random Forest: The Importance Plot from the Random Forest model ranks the most important features impacting Airbnb prices.

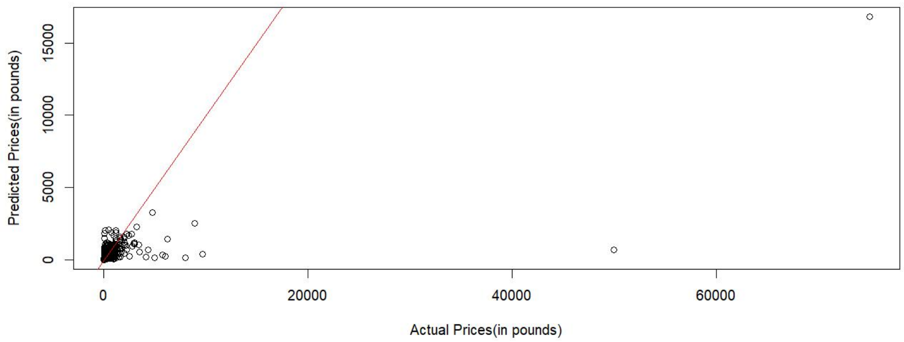
Figure 4: Importance Plot from Random Forest



Distance to Big Ben, Tower of London, Buckingham Palace and British Museum are among the most important features, indicating that proximity to landmarks significantly impacts Airbnb prices. Additionally, number of amenities, neighborhood and host listing count also play an important role, suggesting that both amenities and location are key drivers of price. Airbnb can use this information to help property owners to determine which properties should be listed.

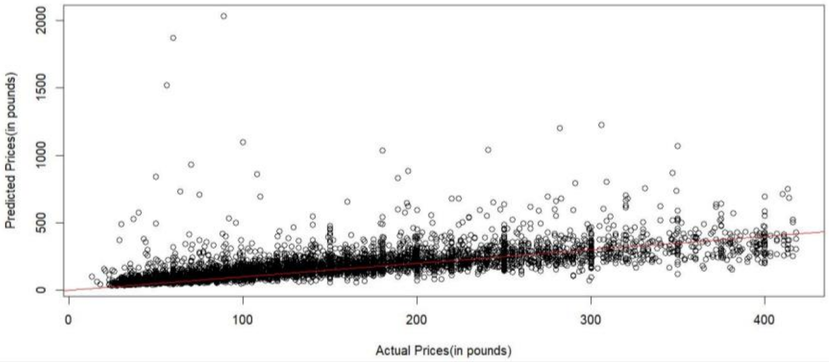
Model performance evaluation: We calculate the R-squared and MAE for our predictions and get MAE of 88 pounds. This MAE is quite high considering the average price for Airbnb listings was around 200 pounds. The scatter plot comparing the actual prices to predicted prices from Random Forest Model gives us more insight into the high MAE.

Figure 5: Model Performance Evaluation



The scatter plot indicates that the Random Forest model reliably predicts prices for most typical listings, but its accuracy decreases for outliers. To evaluate whether removing outliers improves model performance, we excluded them from the test data and reran the predictions, resulting in a MAE of 47.36 pounds. The updated scatter plot is much more interpretable and consistent.

Figure 6: Model Performance Evaluation- removing outliers



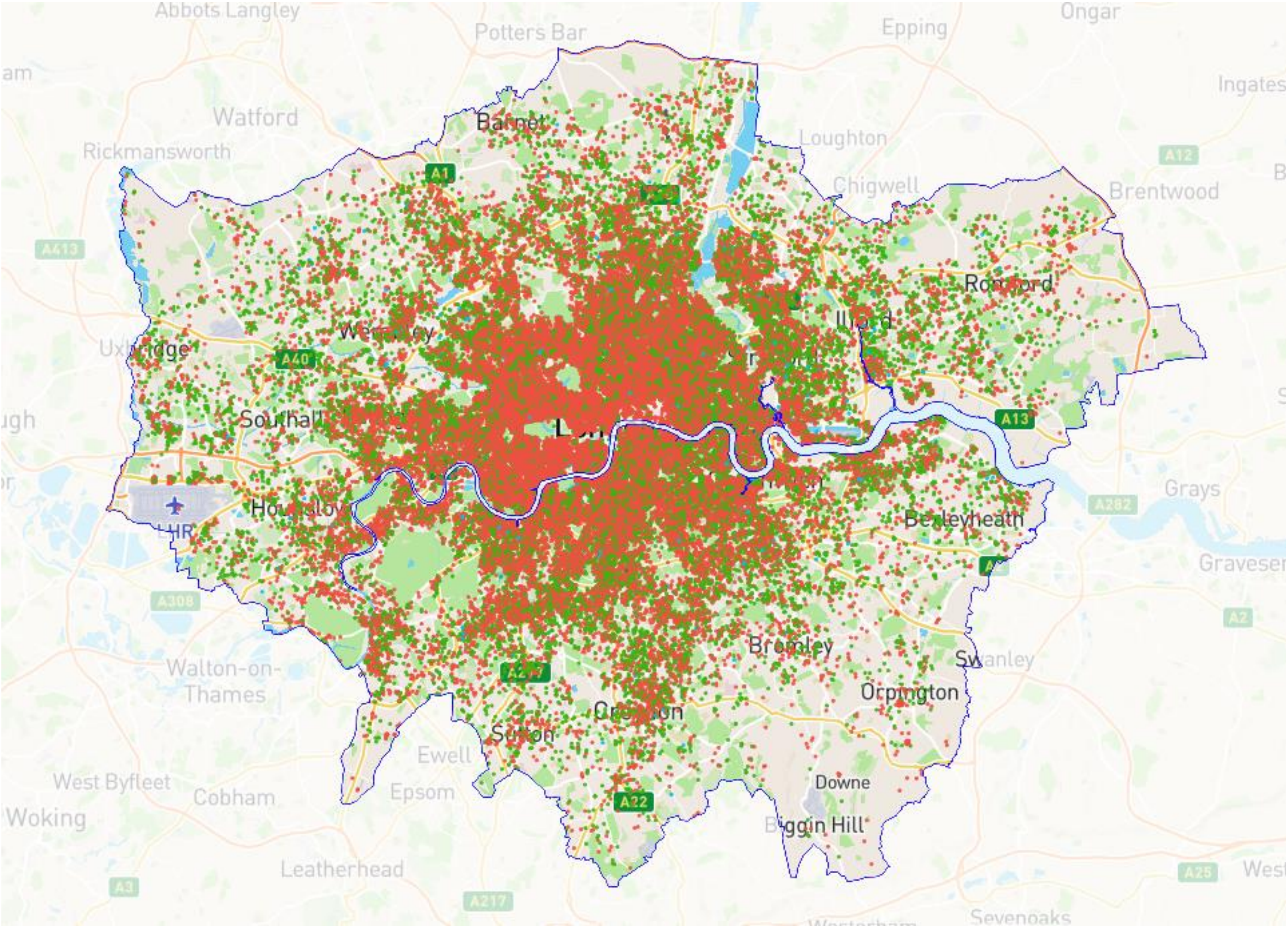
Deployment: The decrease in Mean Absolute Error (MAE) to 47 pounds suggests that flagging and separating outliers during production could enhance our model's predictive accuracy. However, this MAE is still relatively high, indicating a need for either improved data quality or more robust modeling techniques (such as Gradient Boosting, Support Vector Regression, or Neural Networks) to better capture variance in pricing. Consequently, the current model is not ready for deployment. Additionally, we should consider collecting data on occupancy rates, as this information could significantly refine our estimates of earning potential, leading to more accurate and actionable insights for users.

Appendix 1: Details of field data type

Type	**Name of field**
Numeric	id, scrape_id, host_id, host_listings_count, host_total_listings_count, latitude, longitude, accommodates, bathrooms, bedrooms, beds, minimum_nights, maximum_nights, minimum_minimum_nights, maximum_minimum_nights, minimum_maximum_nights, maximum_maximum_nights, minimum_nights_avg_ntm, maximum_nights_avg_ntm, availability_30, availability_60, availability_90, availability_365, number_of_reviews, number_of_reviews_ltm, number_of_reviews_l30d, review_scores_rating, review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, review_scores_value, calculated_host_listings_count, calculated_host_listings_count_entire_homes, calculated_host_listings_count_private_rooms, calculated_host_listings_count_shared_rooms, reviews_per_month
Character	listing_url, source, name, description, neighborhood_overview, picture_url, host_url, host_name, host_location, host_about, host_response_time, host_response_rate, host_acceptance_rate, host_thumbnail_url, host_picture_url, host_neighbourhood, host_verifications, neighbourhood, neighbourhood_cleansed, property_type, room_type, bathrooms_text, amenities, price
Date	last_scraped, host_since, calendar_last_scraped, first_review, last_review

Type	Name of field
Logical	host_is_superhost, host_has_profile_pic, host_identity_verified, neighbourhood_group_cleansed, calendar_updated, has_availability, license, instant_bookable

Appendix 2: Representation Airbnb locations on London map



🟥 Entire home/ apartment 🟩 Private room

Appendix 3: Details of Data Preparation

Step 1: Deleting Fields- Part 1

We removed 36 columns from the initial dataset, each for the following reasons. However, we retained some fields for further transformations and to create more insightful fields in the next phases of the analysis.

- Irrelevant Identifiers: Columns like \`id\`, \`scrape_id\`, \`host_id\`, \`name\`, and \`host_name\` were removed as they are merely identifiers for properties and hosts and do not contribute to predicting the price of Airbnb listings.
- Unnecessary Links: Fields containing URLs such as \`listing_url\`, \`source\`, ‘picture_url’,\`host_thumbnail_url\`, and \`host_picture_url\` were excluded as they are not relevant for price prediction.
- Text Fields Not Suitable for Regression: Character fields like \`description\`, \`host_about\`, \`neighborhood_overview\`, \`host_neighborhood\`and \`bathrooms_text\` were excluded since they are not directly usable for regression analysis. Although we considered quantifying them (e.g., word count), we concluded that such metrics would not significantly impact Airbnb pricing.
- Date and Time Stamps: Fields like \`last_scraped\`, and \`calendar_last_scraped\` were deleted, as they provide time-related information that does not influence property pricing.
- Duplicate and Empty Fields: Columns such as \`host_total_listings_count\`, \`calculated_host_listings_count\`, \`neighborhood\`, \`neighbourhood_group_cleansed\`, \`license\`, \`calendar_updated\` and \`has_availability\` were either duplicates or contained insufficient data to contribute meaningfully to the model.
- Derivative Fields Prone to Multicollinearity: We removed derived fields such as \`minimum_minimum_nights\`, \`maximum_minimum_nights\`, \`minimum_maximum_nights\`,

`maximum_maximum_nights`, `minimum_nights_avg_ntm`, `maximum_nights_avg_ntm`, `calculated_host_listings_count_entire_homes`, `calculated_host_listings_count_private_rooms`, and `calculated_host_listings_count_shared_rooms`. Instead, we retained only core fields like minimum and maximum nights, along with total host listings, to avoid multicollinearity issues.

- Availability Fields (Long-Term): The field `availability_365` was excluded, as most Airbnb bookings do not occur so far in advance. Instead, we kept availability metrics for the next 1, 2, and 3 months, which better reflect short-term property availability. It's worth noting that availability fields are influenced by both customer bookings and host decisions to block the property, indicating the scarcity of a property rather than its demand.

Step 2: Transforming and Creating fields

- Creating Categorical Variables: We transformed key fields such as neighbourhood_cleansed (containing neighborhood information), property_type, and room_type into categorical variables. This allows us to better leverage these features in our regression model by encoding them appropriately.
- Replacing Empty Values: In the bathrooms, bedrooms, and beds columns, we replaced blank cells with zeros. Our analysis indicated that there were no actual zero values in these columns, so empty cells most likely represented properties with zero units.
- Converting Host Since to Years: We transformed the host_since field into a numerical variable representing the number of years the host has been on Airbnb.
- Identifying Listings with Reviews: We created a categorical variable from the review_scores_rating column to identify whether a listing has any reviews. This allows us to analyze the impact of reviews on price.

- Accounting for Amenities: The amenities column, which originally contained a list of amenities separated by commas, was transformed to make it more useful. We created a new variable to count the total number of amenities and added dummy variables for specific amenities that could influence pricing, such as a refrigerator, kitchen, bathtub, Wi-Fi, and dedicated workspace.
- Calculating Distances to Landmarks: Using the latitude and longitude data, we calculated the distance of each Airbnb listing from four major landmarks in London: Big Ben, the British Museum, Buckingham Palace, and the Tower of London. This helps to capture the effect of location on price.

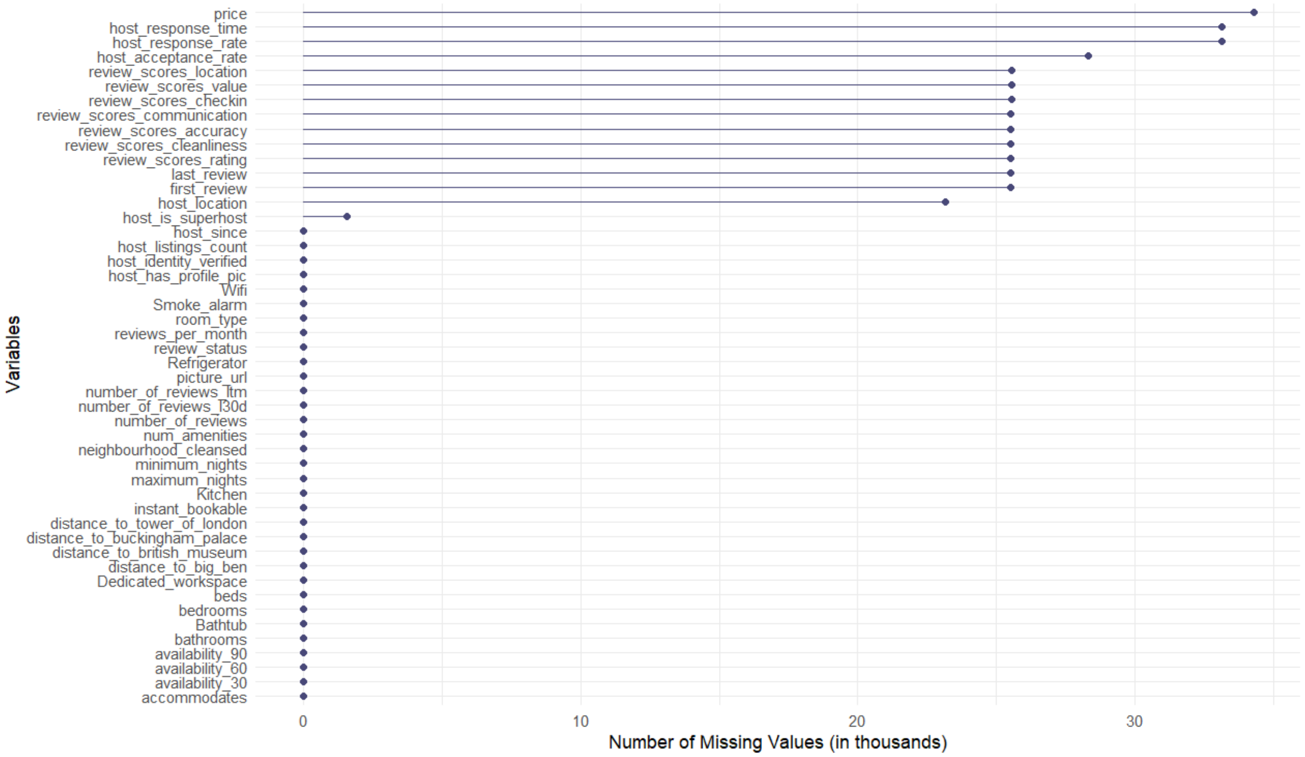
Step 3: Deleting fields- Part 2

- Deleting property_type: We decided to remove the property_type field as it is just a detailed version of the room_type field (e.g. room type- private room, property_type- private room in condo). While property_type provides more detailed information, it contains 98 categories, as revealed by str(airbnb_data). This would present challenges related to overfitting while doing regression. Given its likely minimal contribution to explaining price variance and the complexity it introduces, we opted to exclude this field.
- Removing Redundant Fields: Fields such as amenities, amenities_list, latitude, and longitude were removed since they were used to derive other insightful variables. These fields, having served their purpose in feature engineering, are no longer necessary for the final model.

Step 4: Accounting for missing values

To understand which variables have missing values and the extent of missing values we created the following visualization.

Figure 7: Missing data in Airbnb dataset



The visualization indicates that around 33,000 values are missing from the *price* column, which is our dependent variable. Given that we have a total of 96,183 records, removing the rows with missing *price* values would still leave us with over 60,000 records—ample data for running a strong regression model.

However, several other fields, including *host response*, *host acceptance*, *host location*, and *review scores*, have nearly 30,000 missing values. Importantly, our analysis revealed that there is minimal overlap between the missing values in these fields and those in the *price* column. Removing the rows with missing values from these fields would further reduce the dataset to around 33,000 records, which would leave us with less than half of the original data.

Given the potential insignificance of *host response*, *host acceptance*, and *review scores* in explaining *price*, and the substantial reduction in dataset size that would result from keeping these fields, we decided to remove these columns from the dataset. This allows us to preserve a larger and more manageable dataset without compromising the quality of our analysis. After removing the NA values for price and removing the aforementioned fields we have 62,024 records.

<u>Overview of cleaned data</u>

- Number of rows: 62,024
- Number of columns: 32

Step 5: Creating train and test dataset

To ensure that the training and testing datasets are well-balanced, we first randomly split the dataset into 90% training and 10% testing, using a set seed of 123 for reproducibility. After splitting, we computed the mean values of the dependent variable, ‘price’, in both the training and testing sets. The close similarity between these mean values suggests that the split is representative of the overall data. Additionally, we used the `ggpairs` function to visualize key variables in both datasets. The similarity in the pairwise relationships in both training and testing sets further confirms that the data split is well-balanced, ensuring that any model trained on this data will generalize well during testing.

Figure 9: GGpairs result for train data

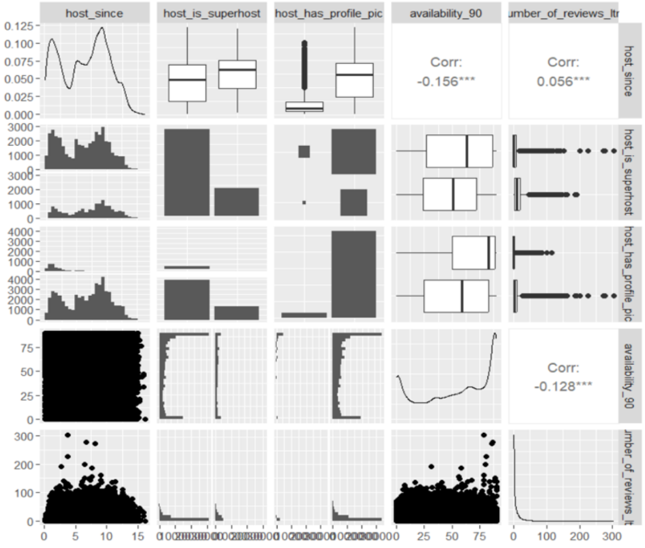
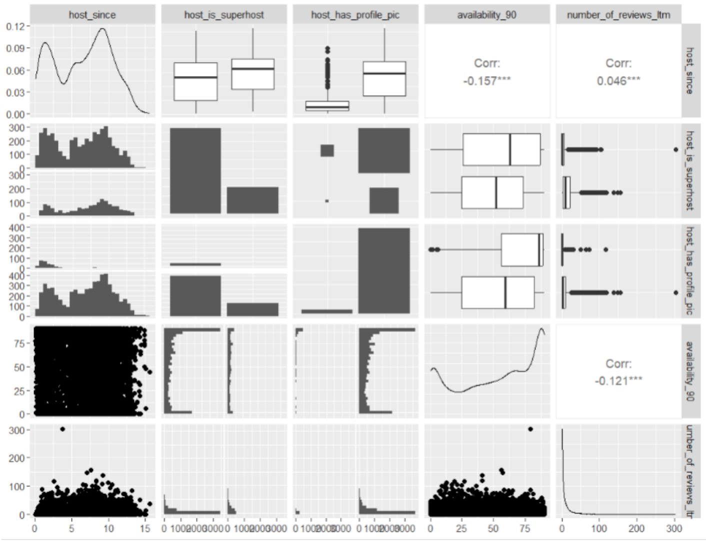


Figure 8: GGpairs result for test data



It is important to highlight that the box-plot visualization of prices in both the training and test datasets reveals the presence of outliers. These outliers may represent extremely high-end properties or errors in price input.

Figure 10: Boxplot of train data

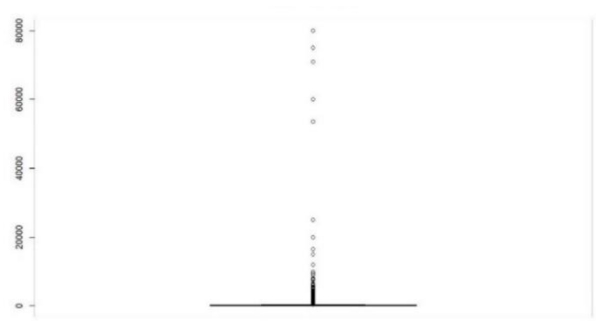
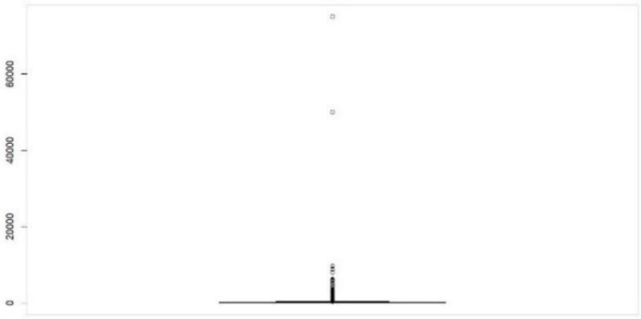
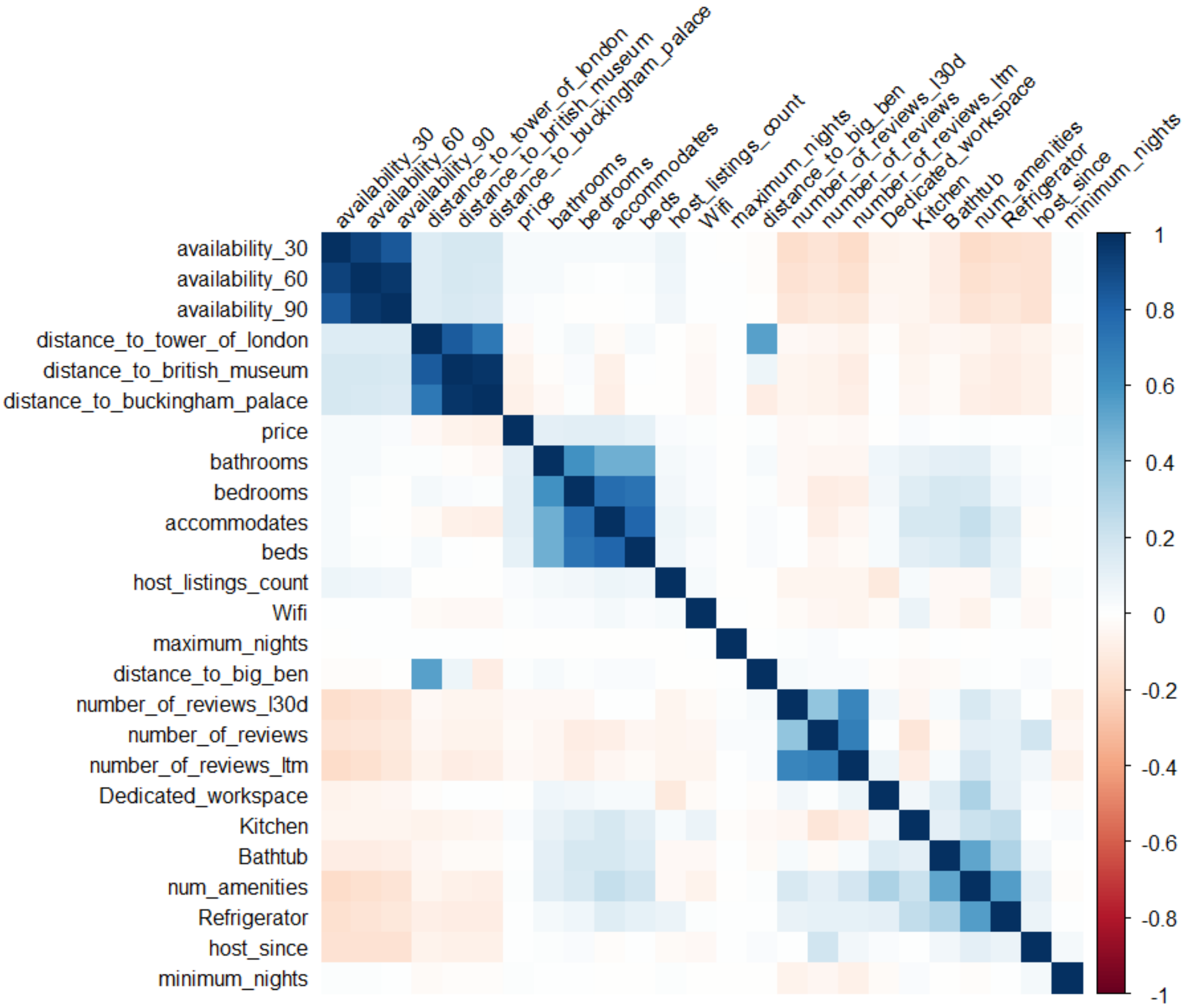


Figure 11: Boxplot of test data



While outliers can pose challenges for some models, we choose to retain them as they reflect real-world scenarios, such as pricing errors and unusually priced properties on Airbnb. By keeping these outliers in our dataset, we can better evaluate which models manage outliers effectively, ultimately improving their predictive accuracy in a production setting.

Appendix 4: Correlation Analysis


availability_30
availability_60
availability_90
distance_to_tower_of_london
distance_to_british_museum
distance_to_buckingham_palace
price
bathrooms
bedrooms
accommodates
beds
host_listings_count
Wifi
maximum_nights
distance_to_big_ben
number_of_reviews_l30d
number_of_reviews
number_of_reviews_ltm
Dedicated_workspace
Kitchen
Bathtub
num_amenities
Refrigerator
host_since
minimum_nights