Capstone Project -2

Advanced Price Prediction Modeling for Airbnb Listings in Seattle, Washington: Leveraging Machine Learning and Model Selection Optimization

Springboard Data Science Career Track

Pramod Acharya

1. Introduction

Airbnb, Inc. is a prominent online marketplace and hospitality service intermediary, having evolved significantly since its inception in 2007. Originally conceived as a simple solution to accommodate guests by placing an air mattress in a living room for a nominal fee, Airbnb has expanded into a global business, boasting an impressive annual revenue exceeding $2.6 billion in 2017. The fundamental premise of Airbnb revolves around hosts utilizing the platform to showcase their properties, offering accommodation services to potential guests. In return, Airbnb garners commissions from each successful booking, solidifying its role as a facilitator in the burgeoning sharing economy.

Crucial to the success of hosts on the platform is the strategic decision-making process involved in setting the prices for their listings. Striking the right balance is imperative, as hosts must avoid pricing their accommodations too high, risking the loss of potential customers, or setting them too low, potentially shortchanging themselves. Optimal pricing involves a nuanced consideration of factors such as property location, features, and competitor pricing to fully capitalize on the true value of the listings and maximize revenue. Despite the availability of Airbnb's listing search feature for reference rates, hosts often find the process time-consuming and challenging, especially when attempting to identify properties with similar features in their vicinity. In response to this, our project aims to employ data analytics to analyze historical listing information comprehensively. This analytical approach will identify key factors that influence pricing decisions, empowering hosts with valuable insights.

Moreover, our project goes beyond mere analysis by incorporating machine learning models. These models will leverage various inputs, including host information, property features, booking policies, and more, to predict listing prices accurately. The specific focus of our endeavor is on the dynamic Airbnb market in Seattle, Washington. Understanding the multifaceted nature of pricing in the Airbnb ecosystem is essential. Factors contributing to price variations include custom pricing set by hosts, low availability, customer demand, and additional charges for guests. It is crucial to acknowledge that listed prices may deviate, and guests often have the opportunity to negotiate discounts directly with hosts, particularly for extended stays or during off-peak seasons.

The primary goal of this data science project is to develop a robust pricing model tailored to Airbnb listings in Seattle, Washington. The inconsistency and variability observed in listing prices, coupled with the lack of clarity on which facilities matter most to guests, drive the need for a predictive model. By analyzing a multitude of facilities or amenities, our model aims to provide guidance for Airbnb hosts in Seattle regarding pricing strategies and future listing plans. Ultimately, this initiative seeks to enhance the overall user experience for both hosts and guests on the Airbnb platform.

This study is undertaken with the overarching objectives of delving into and analyzing Airbnb's listings in Seattle, Washington. The primary goals encompass the identification of features that exert influence on the pricing dynamics of nightly stays and the subsequent development of machine learning models adept at predicting these prices based on pertinent features. The ultimate aim is to furnish hosts with valuable recommendations aimed at optimizing their revenues. The structure of this report is methodically organized into distinct sections. Section 2 meticulously describes the dataset, followed by Section 3, which expounds upon the data cleaning and wrangling process. In Section 4, we engage in an in-depth exploration and statistical analysis of the dataset. The subsequent section, Section 5, is dedicated to the development and evaluation of machine learning models. The final section, Section 6, encapsulates recommendations tailored for Airbnb hosts and provides insightful suggestions for future research endeavors. Through a thorough examination of Toronto's Airbnb market dataset, we endeavor to discern the pivotal features influencing pricing, equipping hosts with a reference point for refining their properties or booking policies. Furthermore, our aim is to construct machine learning models that empower hosts to establish prices that are both equitable and competitive.

2. Data Collection

The dataset utilized for this study was acquired from insideairbnb.com, an activist website dedicated to periodically scraping Airbnb's database to retrieve comprehensive listing information. The specific file selected for analysis is downloaded and subsequently unzipped, resulting in the creation of a file named 'listings.csv.' To access the most recent and relevant data for our examination, we retrieved the dataset from the following link: http://insideairbnb.com/get-the-data/. As of the information provided, the dataset was last scraped on 09/18/2023, ensuring that our analysis is conducted on a current and representative snapshot of the Airbnb listings in Seattle, Washington. This meticulous approach to data collection and the timeliness of the dataset contribute to the accuracy and relevance of our study, enabling a robust exploration and analysis of the factors influencing pricing dynamics in the Airbnb market.

The dataset represents a collection of Airbnb listings in Seattle, Washington, comprising 6,823 rows and 75 features. Each row corresponds to a distinct listing available for rent on the Airbnb platform within the specified location. Below is a brief description of the dataset and the types of data it contains:

2. 1. Property Information

The dataset includes details about each property such as its unique identifier (`id`), name (`name`), description (`description`), type of property (`property_type`), type of room available (`room_type`), number of guests accommodated (`accommodates`), number of bedrooms (`bedrooms`), and number of beds (`beds`).

2.2. Location Information:

Location-related information comprises latitude (`latitude`) and longitude (`longitude`) coordinates, as well as neighborhood details in both cleaned and uncleaned formats (`neighbourhood_cleansed`, `neighbourhood_group_cleansed`).

2.3. Price and Availability:

Pricing information includes the nightly rate (`price`), minimum and maximum nights required for booking (`minimum_nights`, `maximum_nights`), and availability for different time frames (`availability_30`, `availability_60`, `availability_90`, `availability_365`).

2.4. Review Information:

The dataset encompasses review-related data, such as the total number of reviews (`number_of_reviews`) and overall rating score given by guests (`review_scores_rating`).

2.5. Host Information:

Details about the hosts include their unique identifier (`host_id`), name (`host_name`), registration date (`host_since`), location (`host_location`), description (`host_about`), response time (`host_response_time`), response rate (`host_response_rate`), acceptance rate (`host_acceptance_rate`), superhost status (`host_is_superhost`), profile picture availability (`host_has_profile_pic`), and identity verification status (`host_identity_verified`).

2.6. Instant Booking:

The dataset indicates whether instant booking is available for each listing (`instant_bookable`).

2.7. Calculated Host Listings:

Information about the host's listings includes the total count of listings (`calculated_host_listings_count`), count of entire homes (`calculated_host_listings_count_entire_homes`), count of private rooms (`calculated_host_listings_count_private_rooms`), and count of shared rooms (`calculated_host_listings_count_shared_rooms`).

2.8. Policy Information:

The dataset may contain information about the cancellation policy (`cancellation_policy`).

2.9. Airbnb Listing Information:

URLs for the listing (`listing_url`) and associated pictures (`picture_url`) are included.

2.10. Web Scraping Information:

Web scraping details consist of the scrape ID (`scrape_id`) and the date when the listing was last scraped (`last_scraped`).

The dataset encompasses diverse information essential for understanding and analyzing Airbnb listings in Seattle, facilitating various types of analyses, including pricing trends, occupancy rates, host behavior, and customer preferences.

3. Data Cleaning and Wrangling

Data wrangling is an essential preparatory phase for effective analysis of Airbnb listing prices in Seattle, Washington. The overarching goal is to enhance the dataset's quality and utility for subsequent exploratory data analysis (EDA) and statistical examination. One key aspect involves ensuring correct data types for each feature, guaranteeing consistency and accuracy. This entails converting numerical values to appropriate types and addressing any discrepancies in data type assignments. Additionally, the process focuses on handling missing data judiciously, employing strategies like imputation to prevent biased analyses. The creation of potentially useful features is another facet, introducing new variables that may offer deeper insights into pricing dynamics. Finally, data wrangling aims to ready the dataset for EDA by organizing it to facilitate meaningful analyses, addressing outliers, and preparing for statistical examination. These steps collectively ensure that the dataset is well-structured and reliable for uncovering patterns and understanding the factors influencing Airbnb listing prices.

Specifically tailored for Airbnb listing prices in Seattle, data wrangling involves meticulous steps like validating and converting the 'price' column to a numeric format, addressing missing values in critical pricing features, and creating new features such as 'price per bedroom.' Outliers in the 'price' distribution are handled to create a more representative dataset. These actions collectively contribute to a cleaner, more reliable dataset, establishing a solid foundation for meaningful analyses related to pricing trends and the various factors influencing the cost of Airbnb listings in Seattle.

1. Handling Nightly Prices:

Identification of Outliers:

- Identified listings with nightly prices exceeding $5000.
- Explored and recorded details of the high-priced listings.

Data Cleanup:

- Created a new column `actual_price` to record the actual prices for high-priced listings.
- Dropped the potentially redundant column `host_total_listings_count`.

2. Host Information:

Handling Host-related Columns:

- Examined host-related columns like `host_listings_count` and `host_total_listings_count`.
- Dropped `host_total_listings_count` if it was identical to `host_listings_count`.

3. Distribution of Prices by Room Type:

Exploration of Room Types:

- Calculated mean prices for different room types.
- Visualized the average prices using a horizontal bar chart

4. Geographical Information:

Location Data Exploration:

- Investigated latitude and longitude information.

5. Property Information:

Accommodation Details:

- Explored features related to accommodation such as `accommodates`, `bedrooms`, and `beds`.

- Identified and examined properties with extreme values.

6. Booking Policy:

Minimum Nights Analysis:

- Investigated the distribution of minimum nights.

- Addressed an unusual case with a maximum nights value of 999.

7. Availability:

Availability Metrics Exploration:

- Explored metrics related to availability, including `availability_30`, `availability_60`, `availability_90`, and `availability_365`.

8. Reviews:

Review-related Features Exploration:

- Explored features related to reviews, including `number_of_reviews`, `number_of_reviews_ltm`, and various `review_scores`.

9. Missing Values Imputation:

Identification of Missing Values:

- Examined the degree of missingness in each row.

Numeric Features Imputation:

- Imputed missing values in numeric features like `host_response_rate`, `host_acceptance_rate`, `bedrooms`, and `beds`.

Categorical Features Imputation:

-Imputed missing values in categorical features like `description`, `neighborhood_overview`, `host_location`, etc.

10. Datetime Features:

Exploration of Datetime Columns:

- Examined columns with datetime values like `host_since`, `first_review`, and `last_review`.

Datetime Features Imputation:

- Imputed missing values in datetime features like `first_review` and `last_review`.

11. Correlation Analysis:

Key Features Correlation:

- Selected relevant columns and generated a correlation matrix.

- Visualized correlations using a heatmap.

- Identified pairs with correlation coefficients > 0.75 and created scatter plots for these pairs.

12. New Feature Creation:

Amenities Feature Creation:

- Created a new column `amenities_list` by converting the `amenities` column into a list.

- Generated a set of unique amenities for exploration.

4. Exploratory Data Analysis

The analysis of nightly prices in the 'clean_df' dataset, comprising 6822 entries, indicates a right-skewed distribution with a mean price of $193.97 and significant variability (standard deviation of $275.43). Outliers, notably a maximum price of $10,000, suggest the need for log transformation for more accurate analysis. Descriptive statistics show a range of $9987 and an interquartile range (IQR) of $112. No null values were found in the 'price' column. Log transformation was applied, and visual representations (histogram and box plot for 'log_price') confirm the distributional changes. The average nightly price is approximately $194, with a middle spread of $112 (IQR). Further scrutiny of outliers, especially the maximum value, is recommended for data integrity.

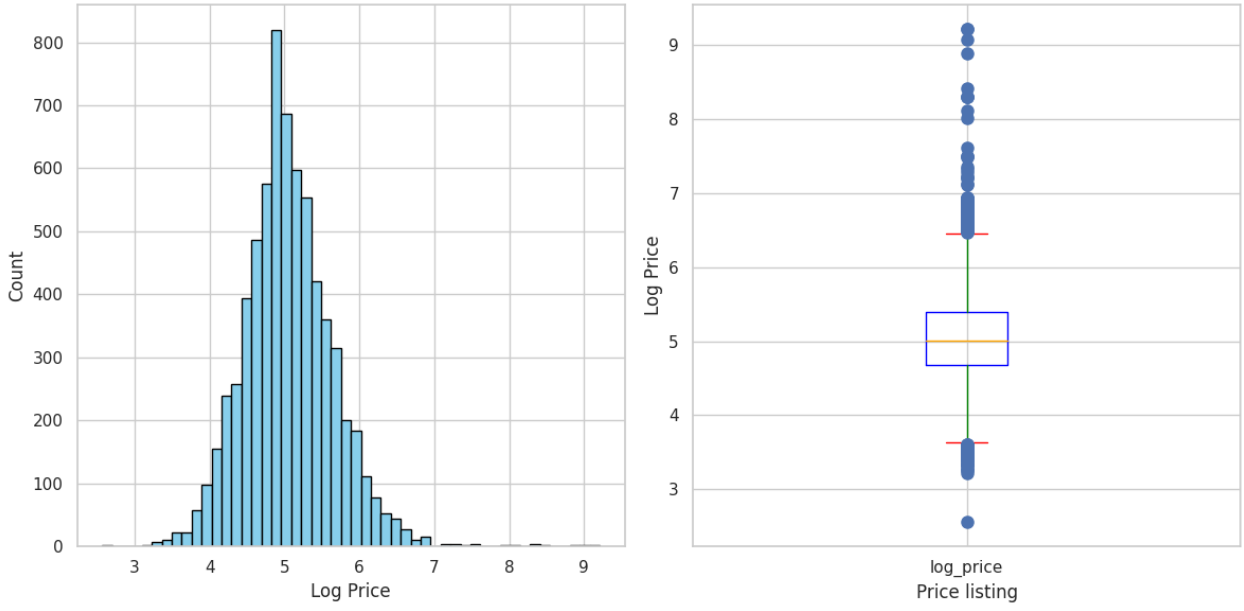
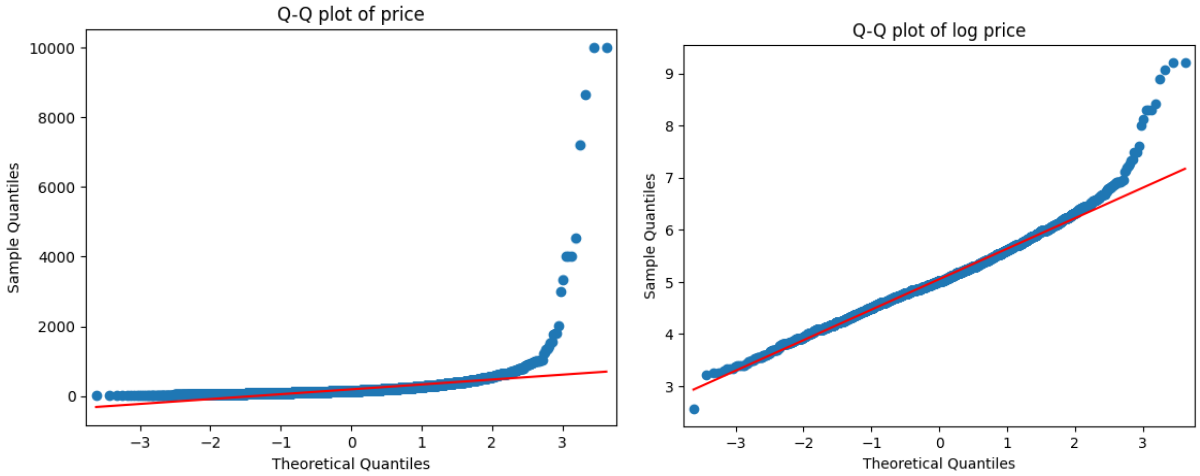


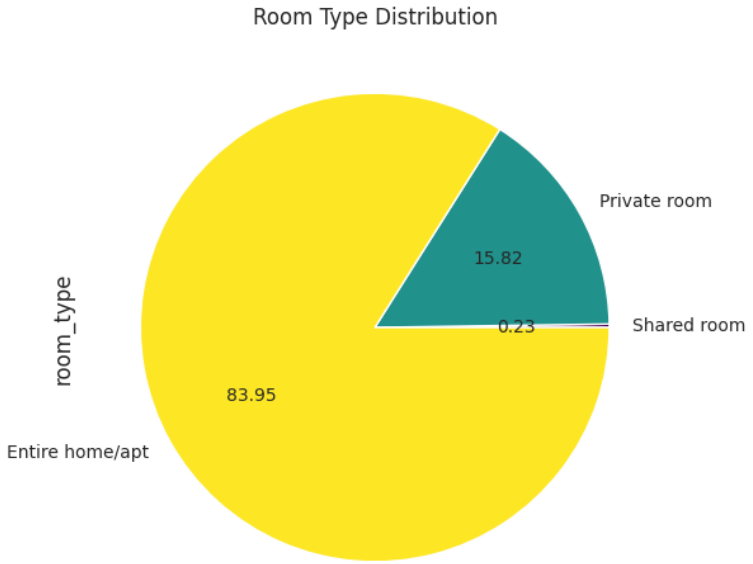
Fig:

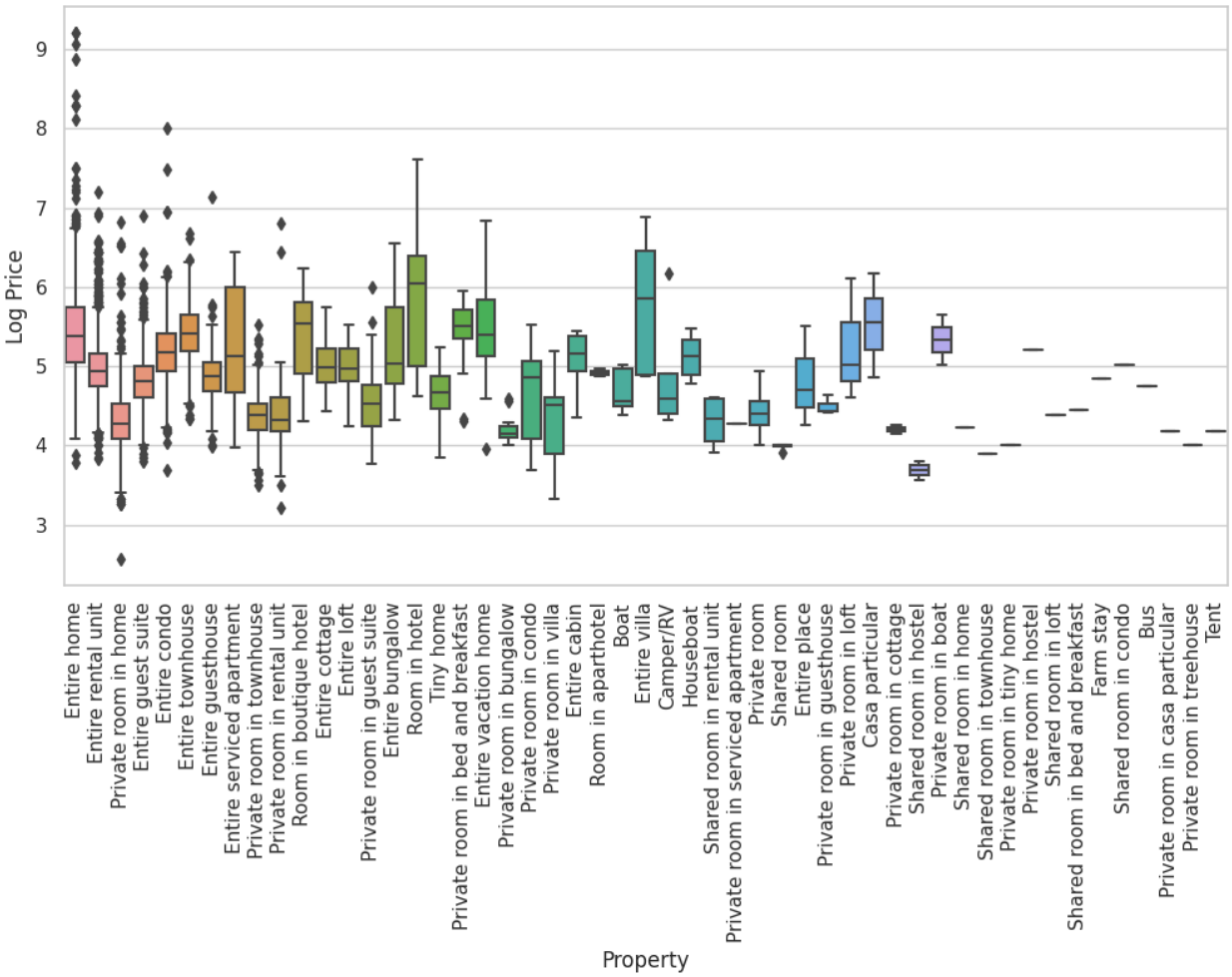
The log transformation of the price variable is implemented to address its right-skewed distribution. Although the log-transformed data retains some skewness, it exhibits a notable improvement toward normality compared to the original distribution. Consequently, the log-transformed variable, `log_price`, will be employed as the target variable for enhanced modeling performance.



4.1 Property Information

In an effort to simplify the analysis and enhance interpretability, the focus shifted to property types with a substantial count of listings, setting a threshold at 3% of the total number of listings. This approach led to the identification of specific property types exceeding the threshold, forming the basis for a more refined examination. The resulting dataset, named `df_top_property_type`, was tailored to capture relevant information, including 'property_type,' 'log_price,' and 'room_type.' A subsequent Seaborn boxplot visualization was employed to explore the relationship between these selected high-count property types and listing prices, revealing potential variations and trends. This strategic refinement addresses challenges associated with a detailed level of granularity, offering a clearer understanding of how specific property types influence pricing dynamics within the dataset.


Room Type Distribution
room_type
Private room
15.82
Shared room
0.23
83.95
Entire home/apt


Log Price
9
8
7
6
5
4
3
Entire home
Entire rental unit
Private room in home
Entire guest suite
Entire condo
Entire townhouse
Entire guesthouse
Entire serviced apartment
Private room in townhouse
Private room in rental unit
Room in boutique hotel
Entire cottage
Entire loft
Private room in guest suite
Entire bungalow
Room in hotel
Tiny home
Private room in bed and breakfast
Entire vacation home
Private room in bungalow
Private room in condo
Private room in villa
Entire cabin
Room in aparthotel
Boat
Entire villa
Camper/RV
Houseboat
Shared room in rental unit
Private room in serviced apartment
Private room
Shared room
Entire place
Private room in guesthouse
Private room in loft
Casa particular
Private room in cottage
Shared room in hostel
Private room in boat
Shared room in home
Shared room in townhouse
Private room in tiny home
Private room in hostel
Shared room in loft
Shared room in bed and breakfast
Farm stay
Shared room in condo
Bus
Private room in casa particular
Private room in treehouse
Tent
Property

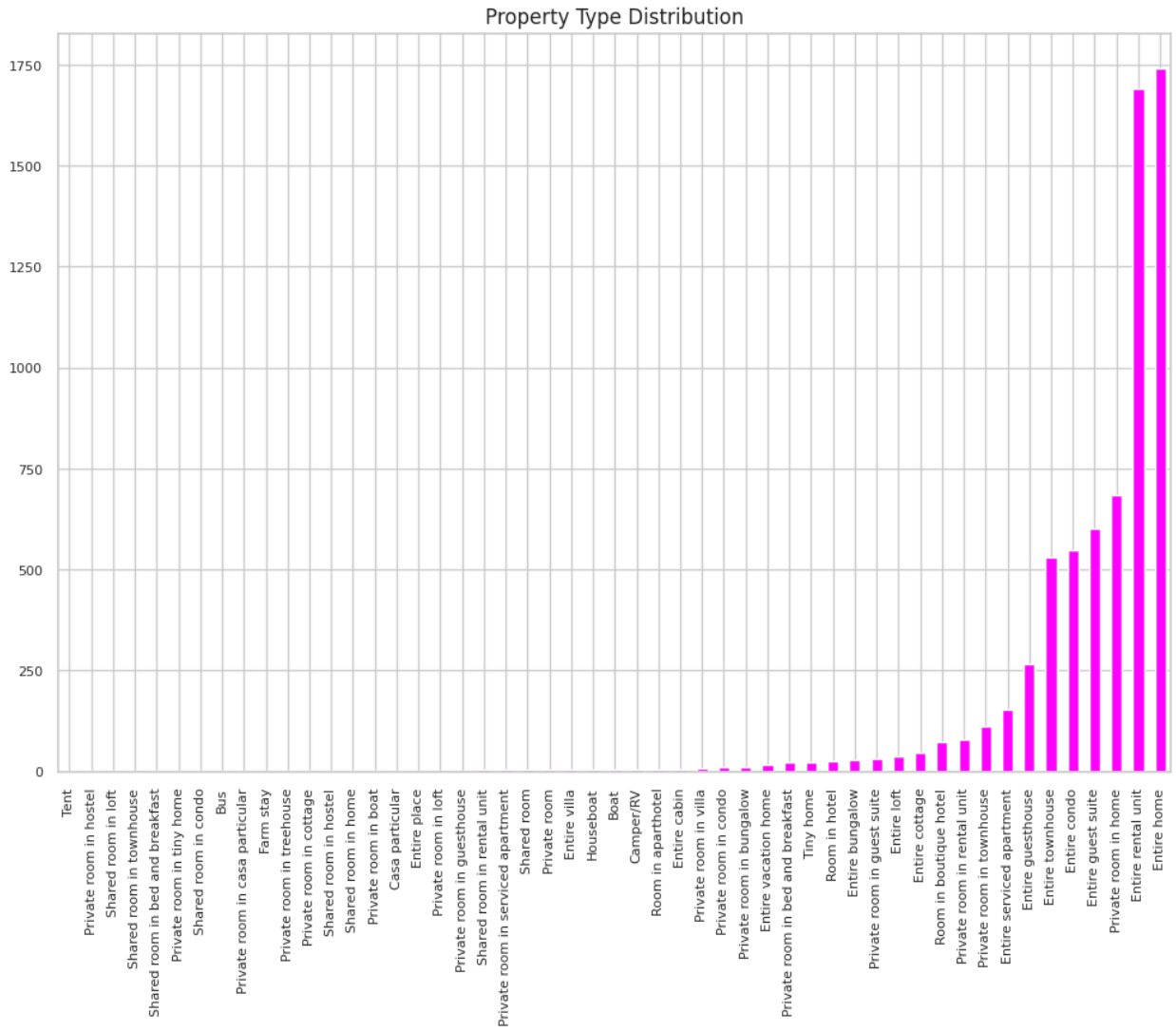
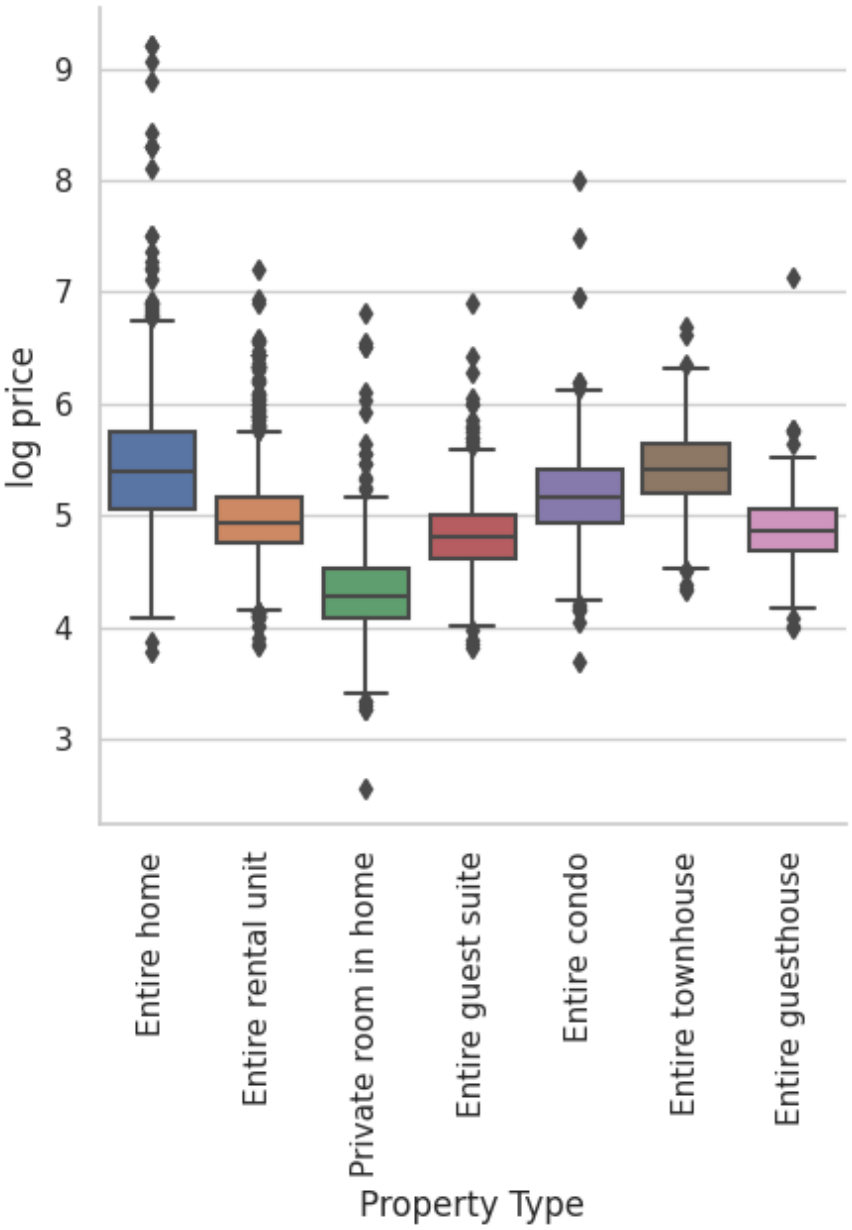


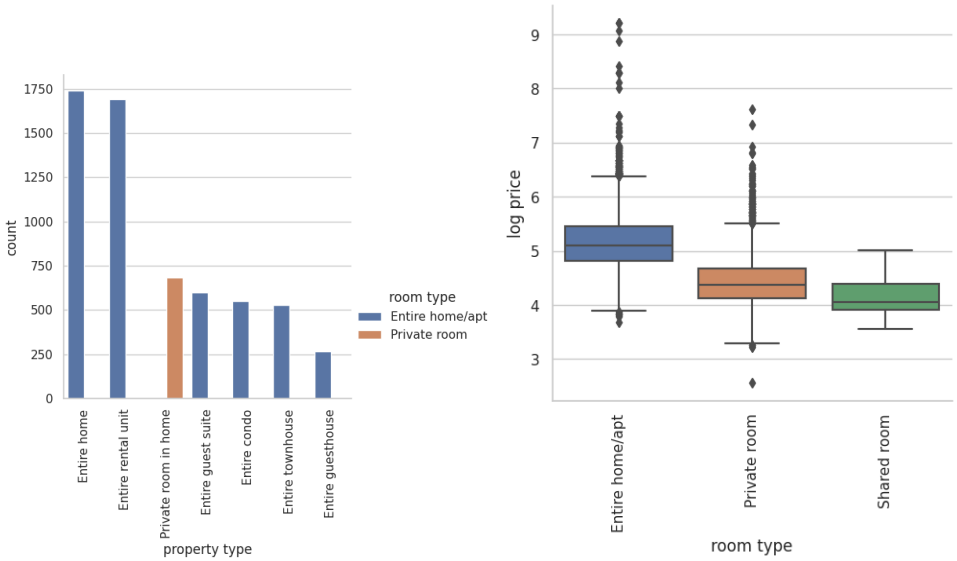
Fig: It appears that entire residences make up roughly 1,700, which is about twice the count of private rooms, approximately 700. Unsurprisingly, apartments and houses dominate the listing count, emphasizing the importance of property and room types in predicting listing prices.

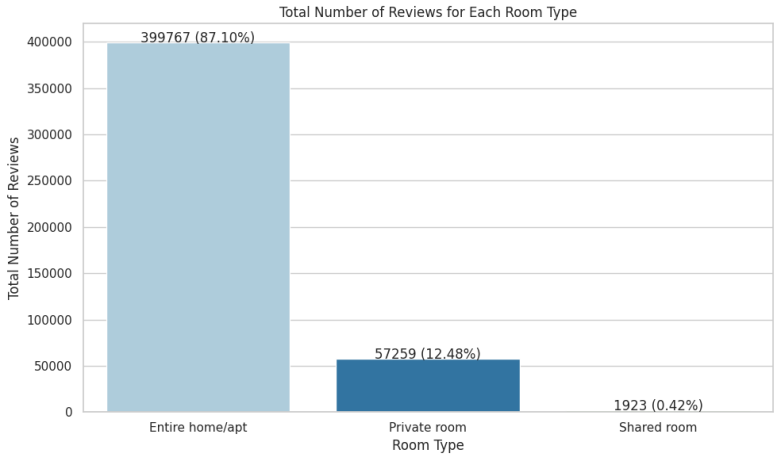


4.2. Available room types

The analysis of room types in relation to pricing is depicted through a Seaborn boxplot. The x-axis represents the various room types, including 'Entire home/apt,' 'Private room,' and 'Shared room,' while the y-axis illustrates the logarithmically transformed listing prices ('log_price'). The resulting visualization allows for a comparative examination of how different room types influence listing prices. Notably, 'Entire home/apt' and 'Entire rental unit' are expected to have a significant impact on pricing, given their prevalence. The boxplot provides insights into the central tendency, spread,

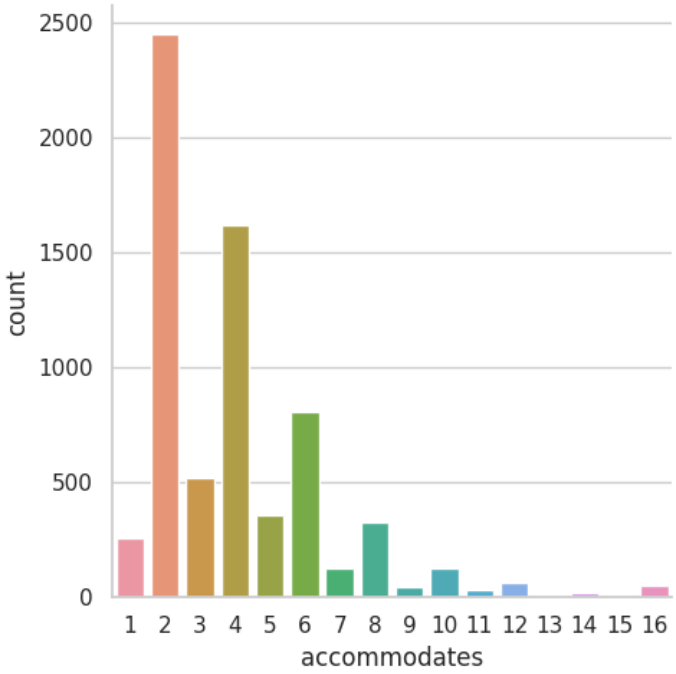
and potential outliers within each room type, contributing to a nuanced understanding of how accommodation type correlates with pricing in the dataset.





4.3. Distribution of number of accommodates

The regression analysis indicates a positive correlation between logarithmic price and accommodation capacity. The left graph suggests a non-linear increase, while the right graph, with a second-order regression line, shows a gradual rise in logarithmic price up to around 12 accommodates, beyond which it plateaus.



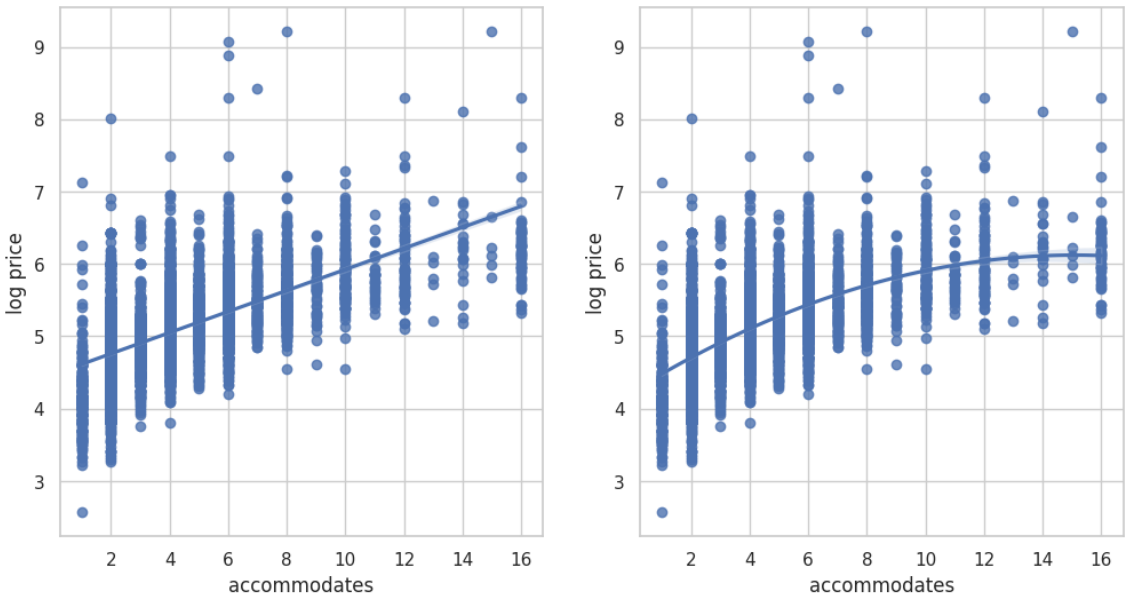


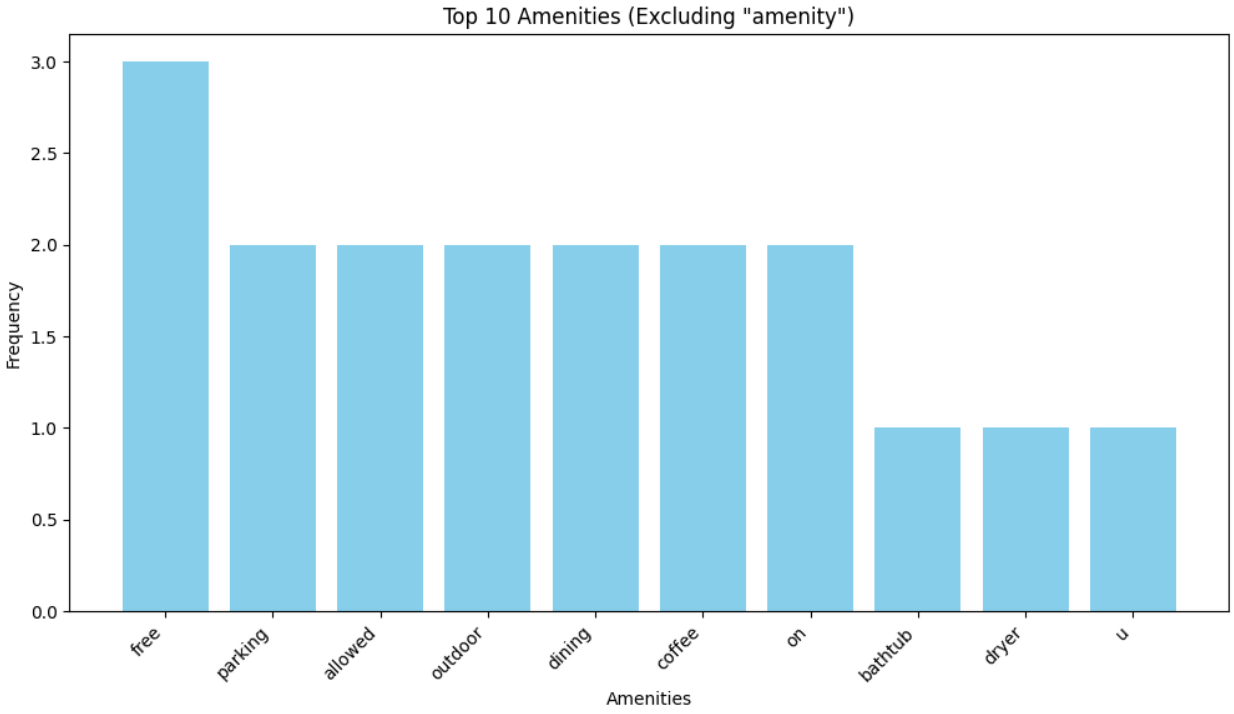
Fig: The logarithmic price correlates positively with accommodation capacity, showing a non-linear increase in the left graph. The right graph, with a second-order regression line, indicates a gradual rise in logarithmic price up to around 12 accommodates, beyond which it plateaus.

4.4. Amenities

The top 20 common amenities include essentials, kitchen, wifi, and others, while the least common 20 include specialized offerings like fast wifi and specific appliances. In terms of price impact, the top 10 amenities leading to the largest change in median price are identified, shedding light on influential features affecting pricing dynamics within the dataset. Regarding amenities, the 20 most common features are identified, providing insights into prevalent offerings among listed accommodations. Understanding these popular amenities is crucial for assessing their impact on pricing and gauging consumer preferences in the market.

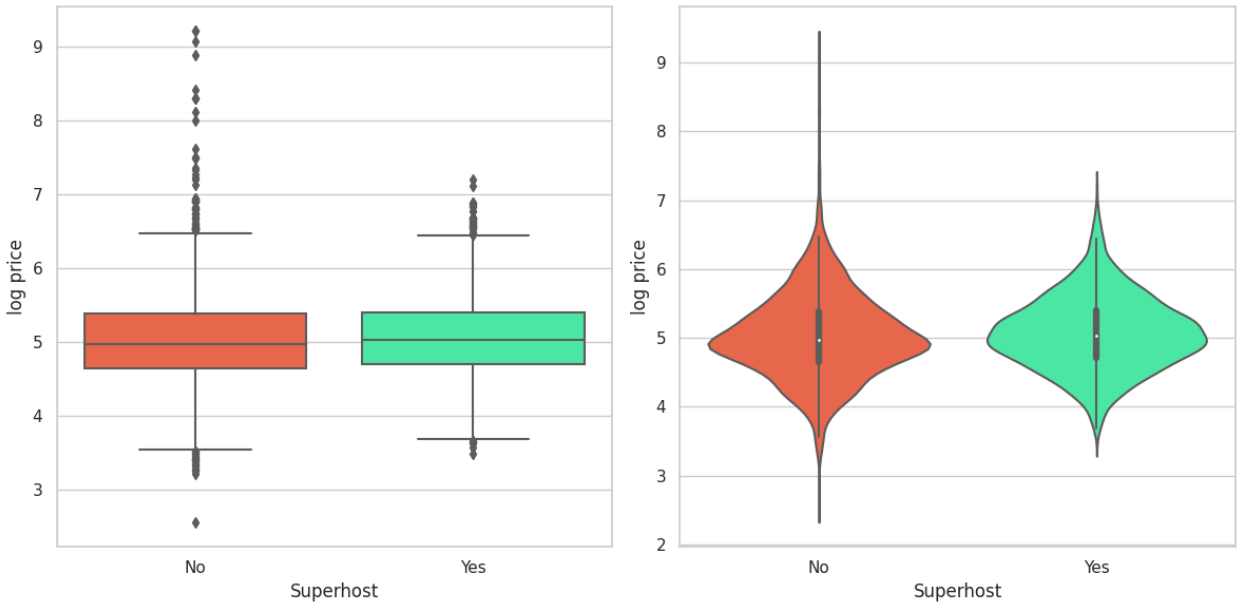
The code performs text processing on a list of amenities, including tokenization, lemmatization, and counting the frequency of each amenity. The resulting bar chart illustrates the top 10 amenities, excluding the term "amenity," based on their frequency in the dataset. The chart provides a visual representation of the most common amenities, offering insights into their prevalence in the Airbnb listings. The code generates a word cloud from lemmatized amenities, providing a visual

representation of their frequency. Larger and bolder words indicate more prevalent amenities, offering a quick overview of popular features in the dataset.

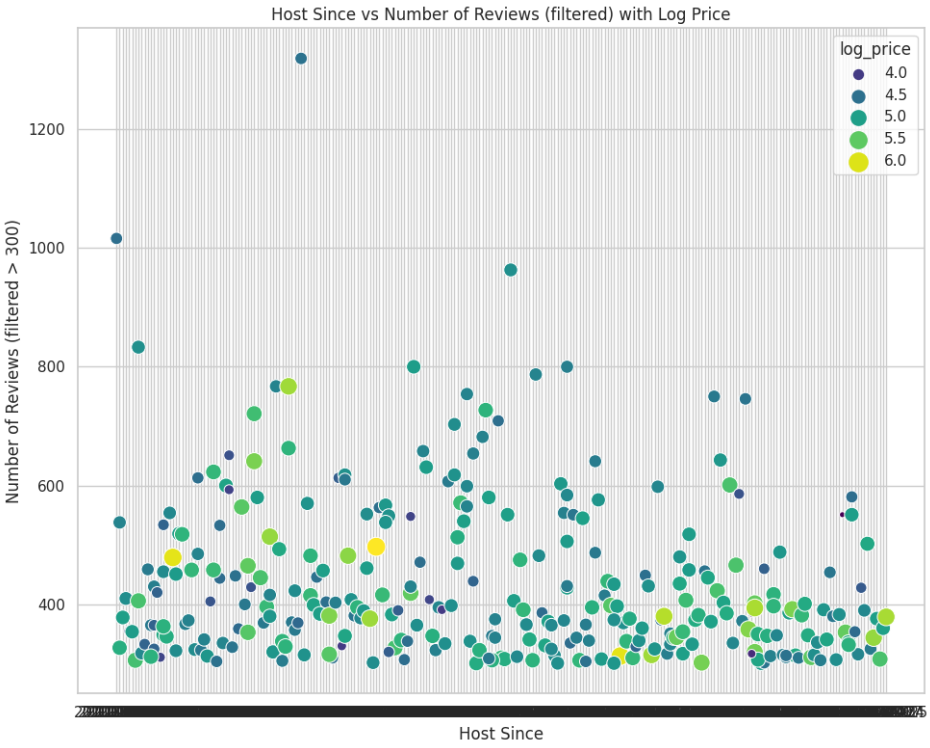


4.5. Host Information

About 46% of listings are from superhosts, known for exceptional Airbnb service. It's important to note a single host can offer multiple listings. Boxplots and violinplots compare the log prices between superhosts and non-superhosts. The visuals suggest similar median log prices, with some upper quartile variations. The violinplot details a comparable overall pattern, indicating a slightly broader range of listing prices for superhosts. These visualizations offer insights into potential pricing differences between the two host categories.

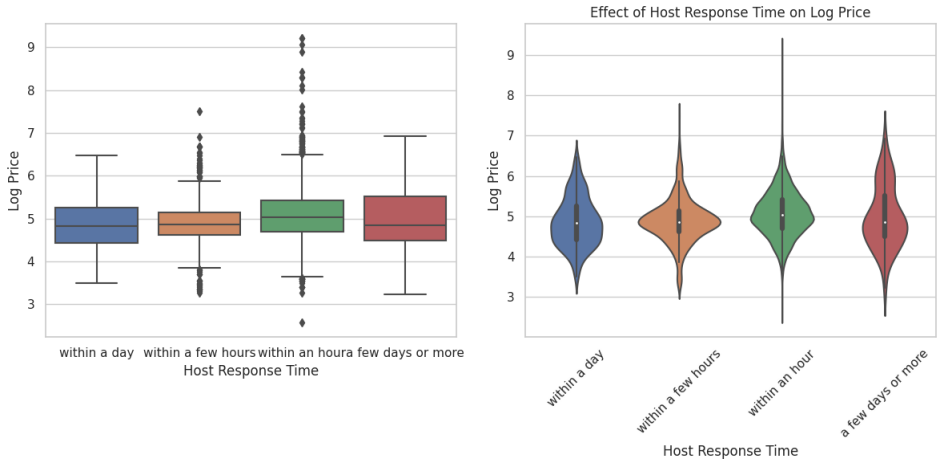


The scatter plot showcases the relationship between the duration a host has been active ('Host Since'), the number of reviews received, and the logarithmically transformed prices. Specifically, the plot focuses on data points where the number of reviews exceeds 300. Each point is sized and colored based on the logarithmic price, providing a comprehensive visual representation. This analysis offers insights into the correlation between host longevity, review count, and pricing dynamics for properties with a substantial number of reviews.



4.6. Host Response Time

Close to 85% of listings have hosts responding within an hour, emphasizing their commitment to prompt communication. Boxplots and violinplots assessing host response time's impact on pricing suggest overall consistency in median log prices across different response times. These visualizations provide insights into the potential relationship between host response time and listing prices.



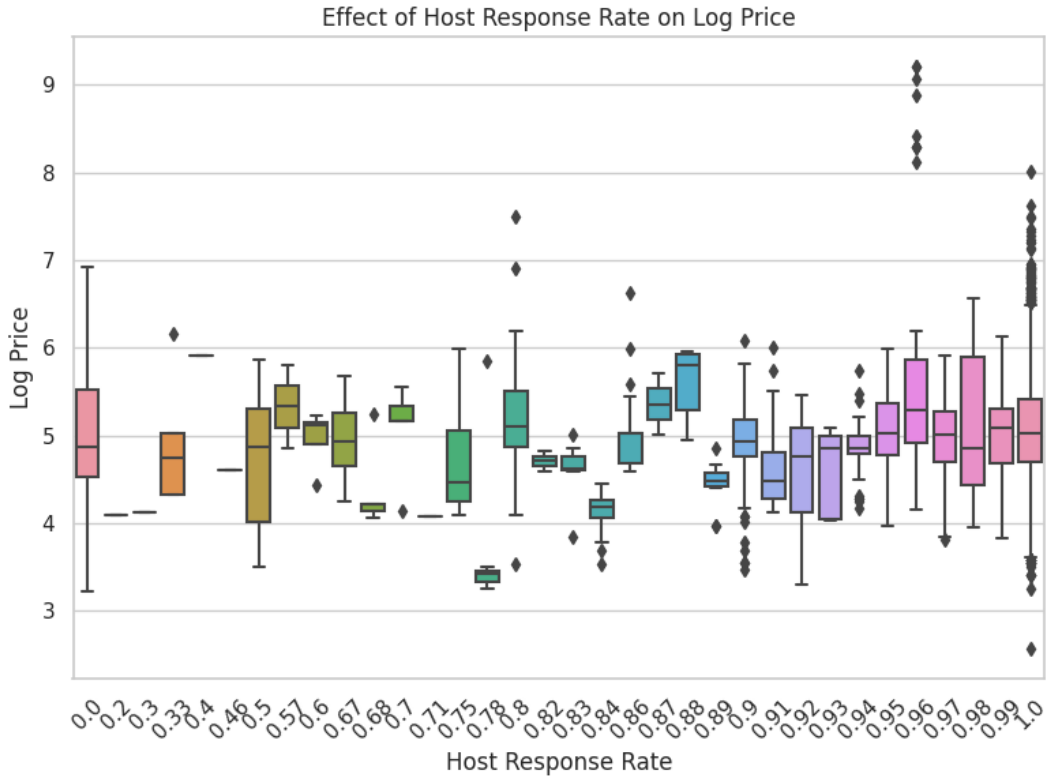
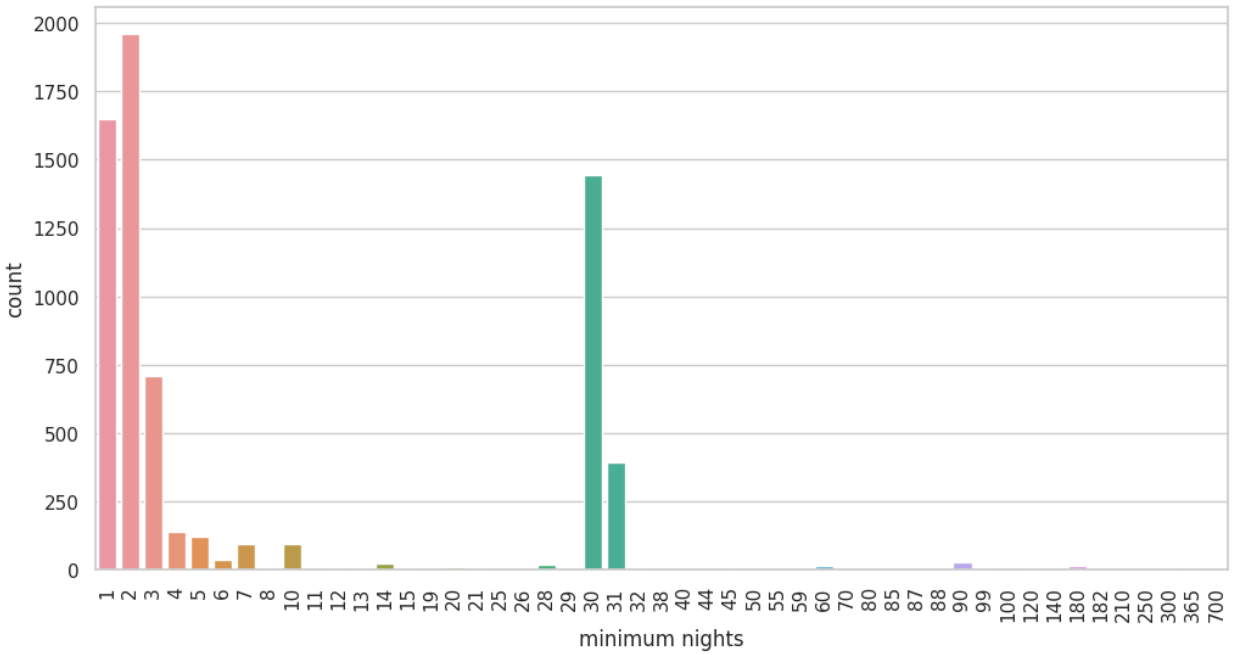
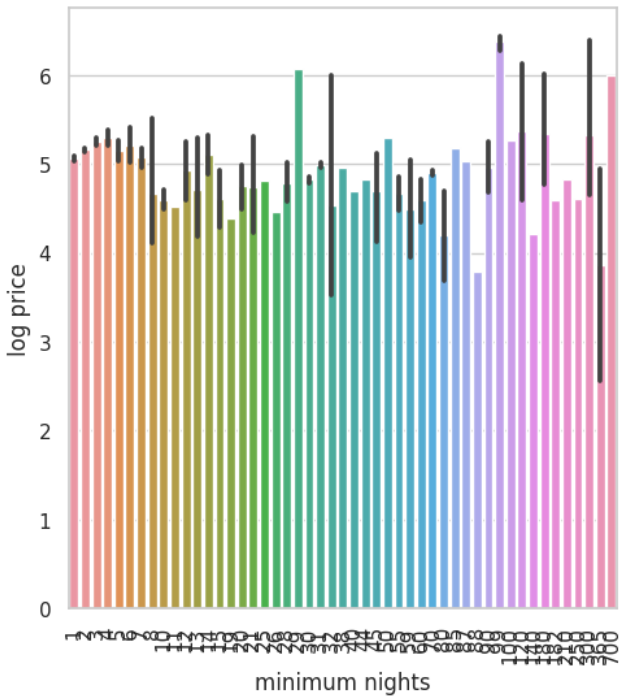


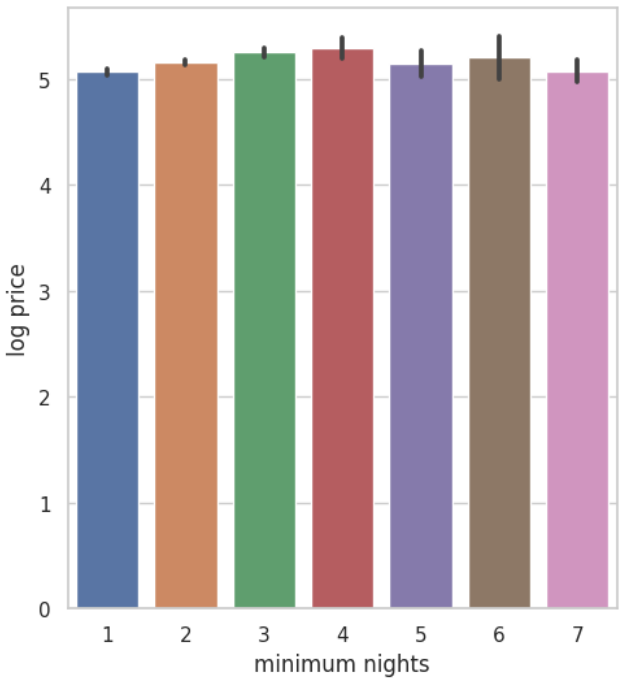
Fig: A boxplot depicts the relationship between host response rates and log prices. The figure, sized 8 by 6, illustrates the potential impact of host response rates on listing prices, offering insights into patterns and variations within the dataset.

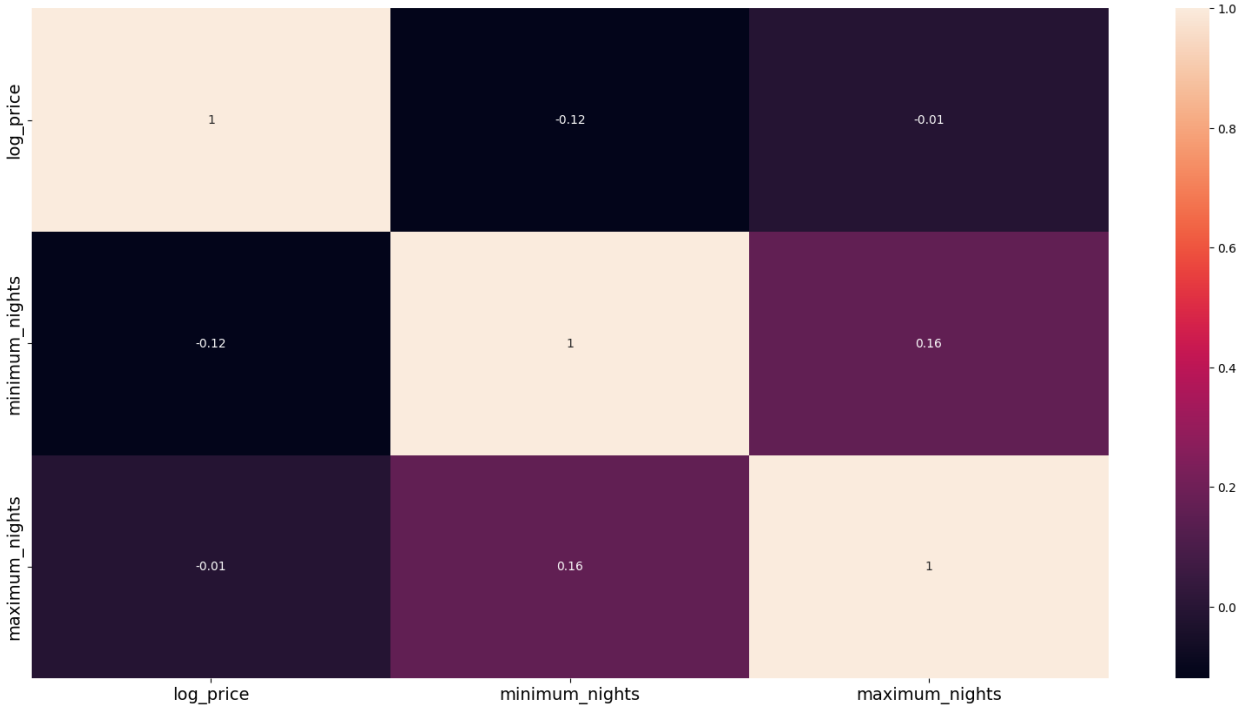
4.7. Minimum Night Stay

The distribution of minimum nights shows a mean of around 11.64 nights, with 69% requiring a stay of 7 days or less. Barplots depict the relationship between minimum nights and log prices. The left plot shows no clear pattern, noting limited data for longer stays. On the right, a distinct median log price difference appears for 1 and 7-night stays, but the relationship isn't linear. These visuals offer insights into the impact of minimum nights on pricing within the dataset.


count
minimum nights

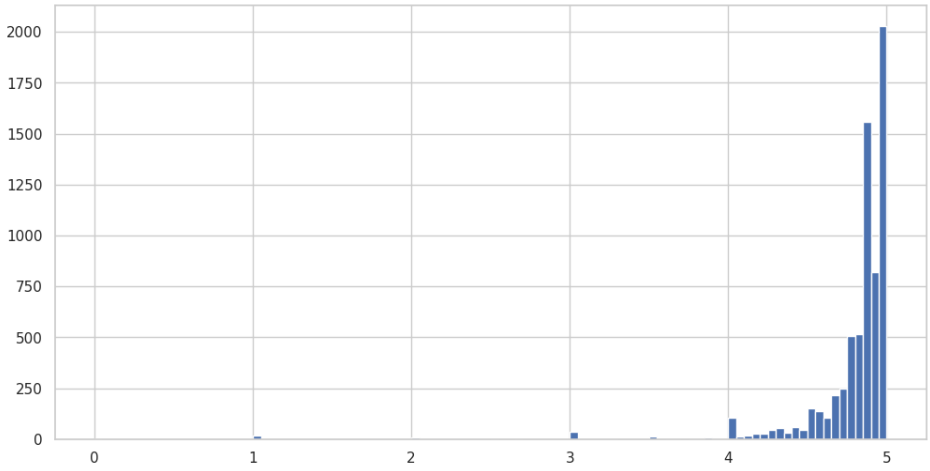

log price
minimum nights

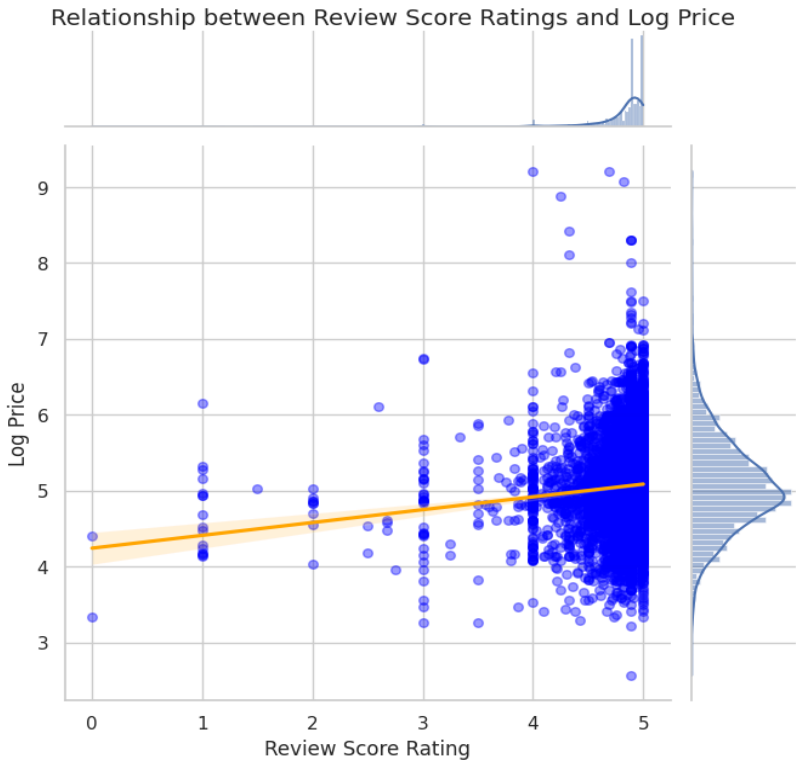

log price
minimum nights



4.8. Review score Rating

Review score ratings exhibit a top-heavy distribution, with a mean of approximately 4.80 and 75% of scores at 4.5 or higher. The range spans from a minimum of 0 to a maximum of 5, showcasing a skew towards positive ratings. Notably, non-superhost listings display more variability, occasionally resulting in lower scores compared to superhost properties.





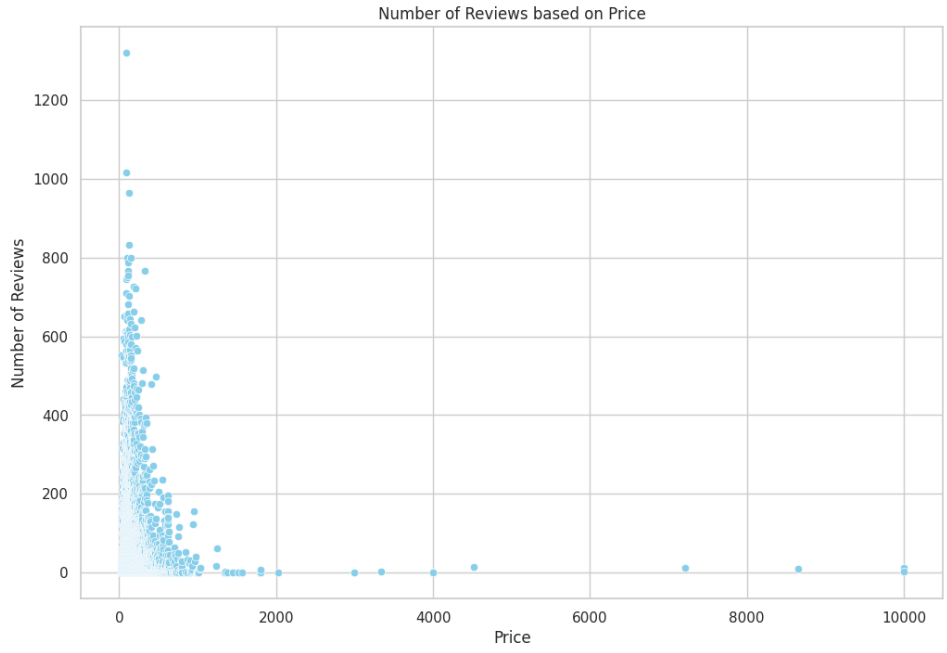
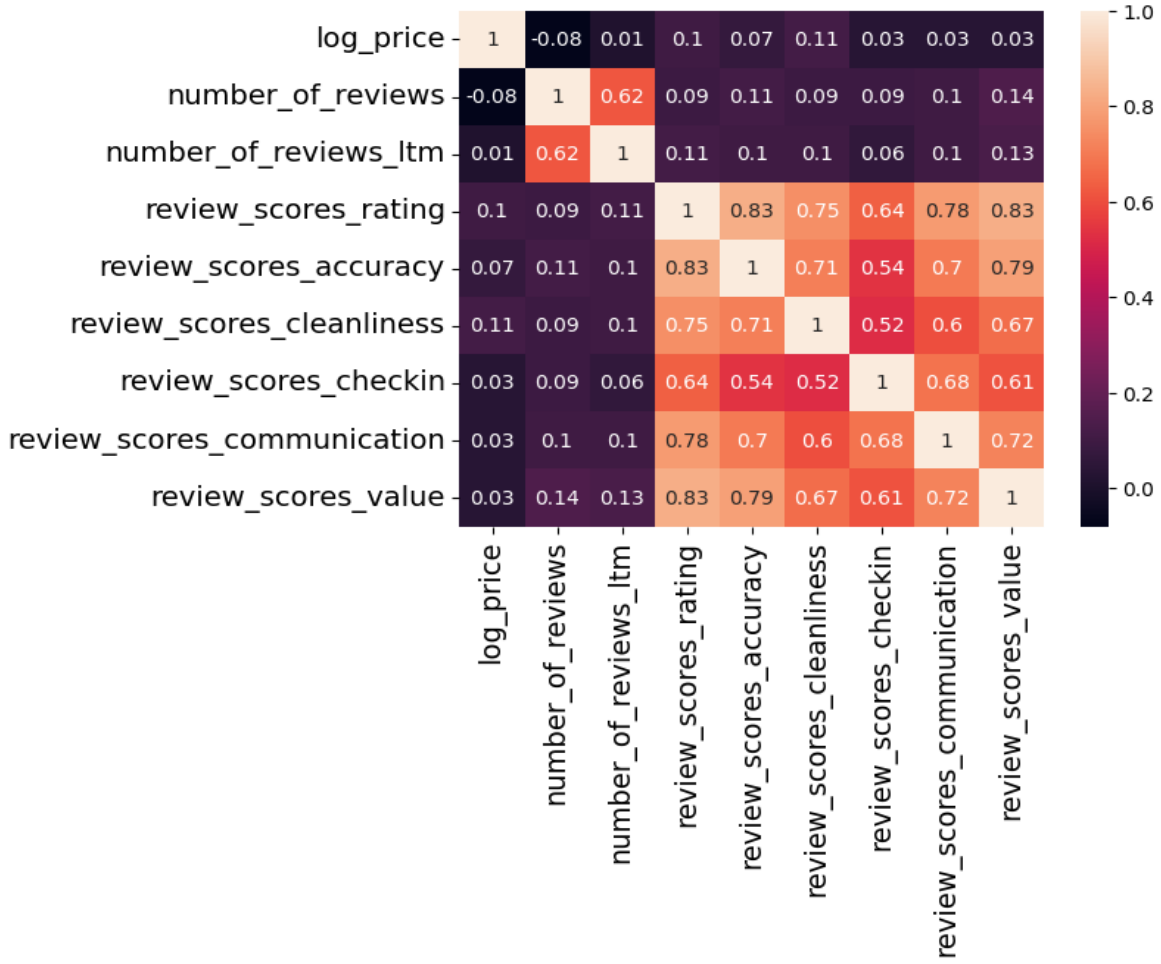
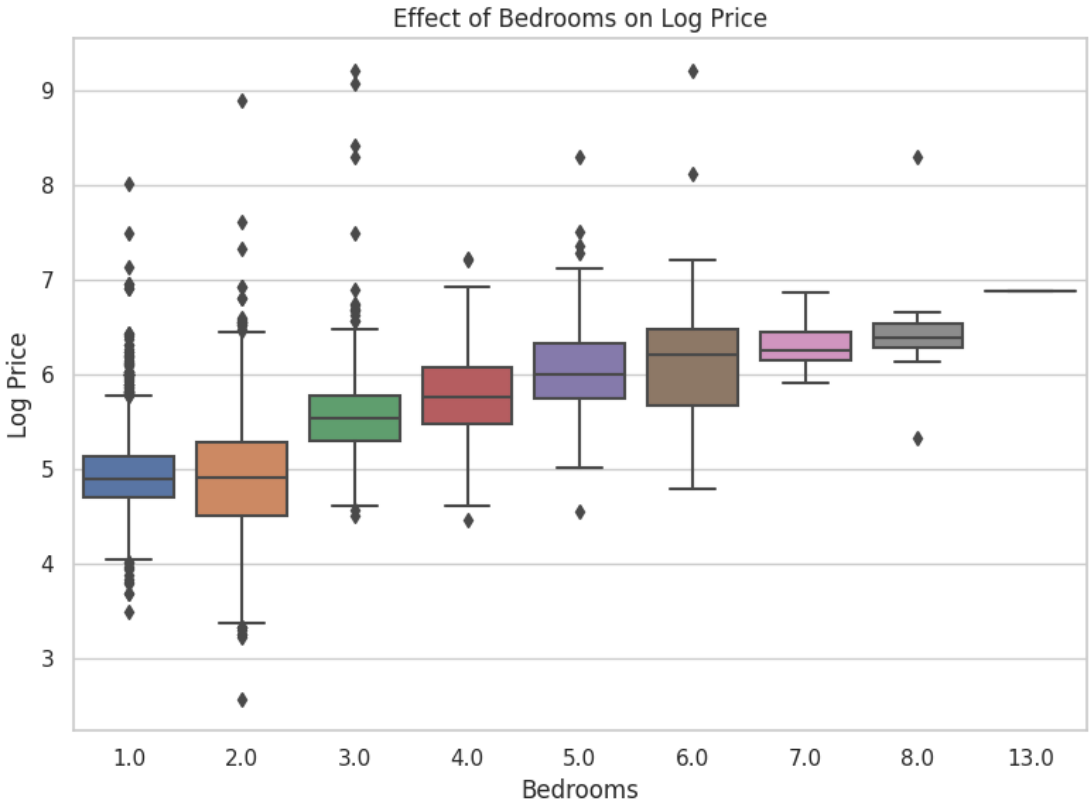


Fig: The scatter plot visually depicts the relationship between listing prices and the number of reviews. Points are color-coded and sorted based on prices, offering insights into how reviews are distributed across different price ranges within the dataset.



4.9. Bedrooms and Bed

The scatter plot visually highlights the positive correlation between the number of bedrooms and median log prices. Each point, representing a specific bedroom count, is shown in blue, while the orange regression line emphasizes the overall upward trend. The accompanying equation quantifies this relationship, providing insights into how changes in bedroom count relate to variations in median log prices. This concise analysis offers valuable insights into the pricing dynamics associated with different bedroom counts in the dataset.



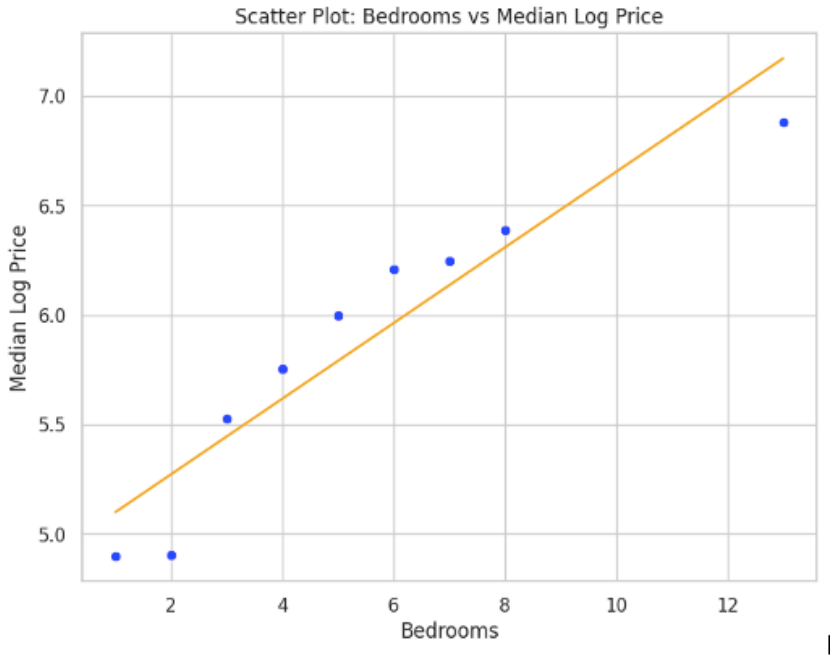
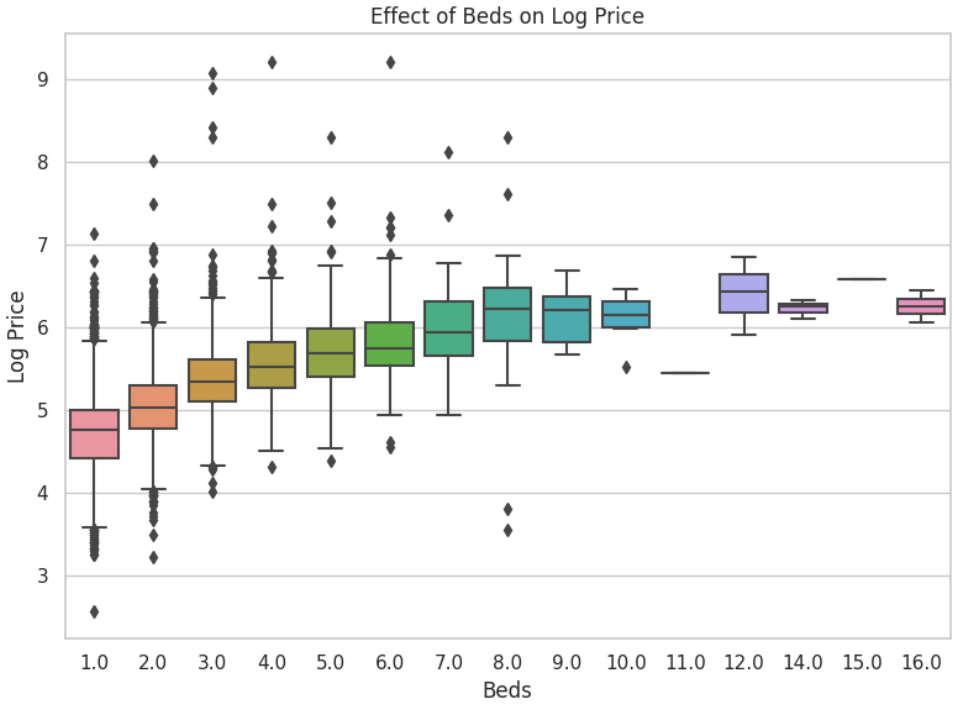


Fig: Linear Fitting y=0.17x+4.93



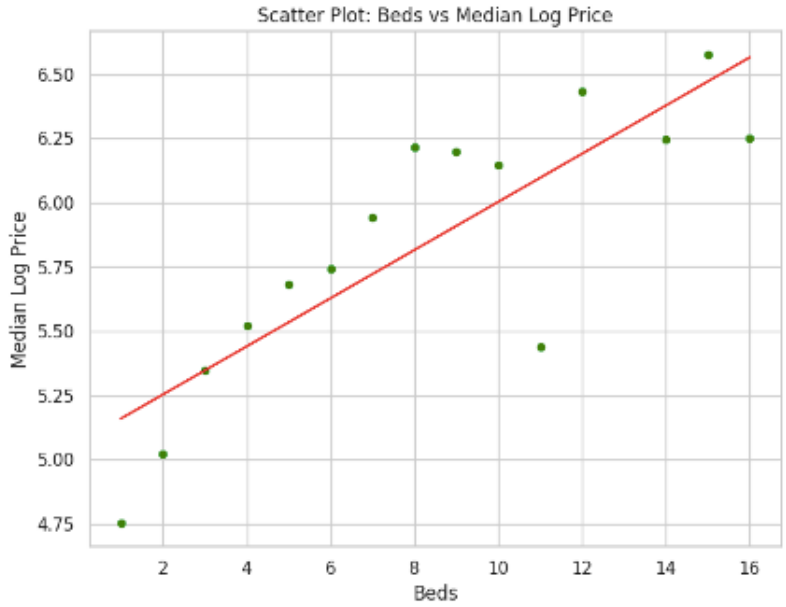
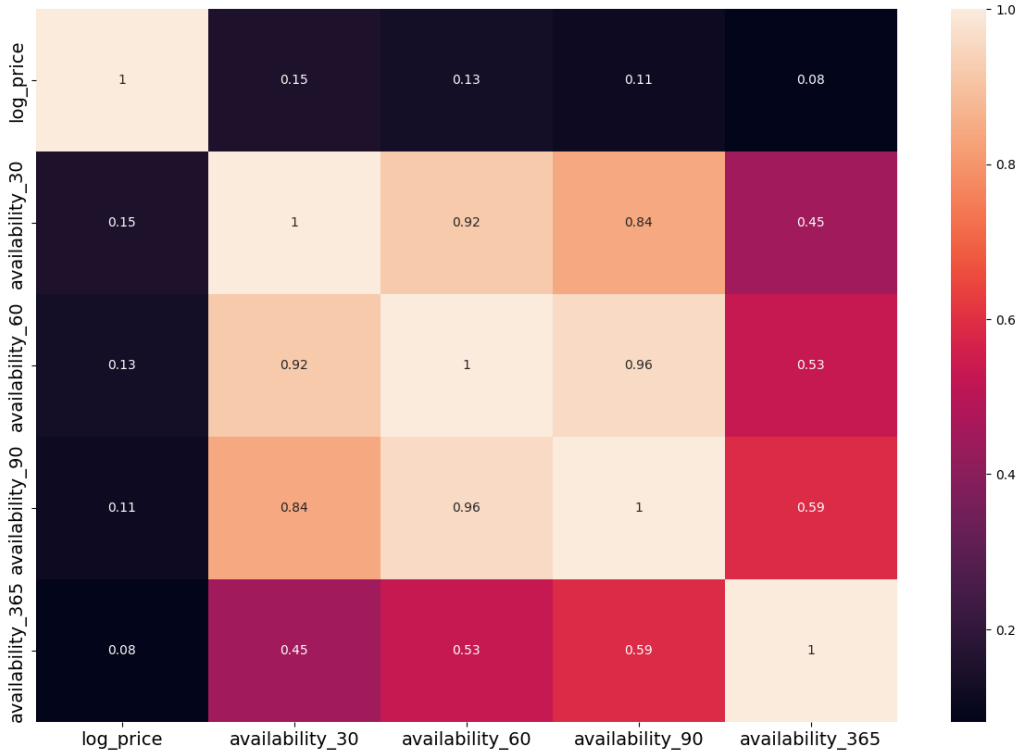


Fig: Linear Fitting: y=0.09x+5.07

4.10. Availability

The heatmap illustrates the correlation between log price and availability metrics, including availability for 30, 60, 90, and 365 days. Notably, log price exhibits no significant correlation with availability. Conversely, high correlations are observed among the availability metrics for different day periods, suggesting a strong interdependence in their patterns.

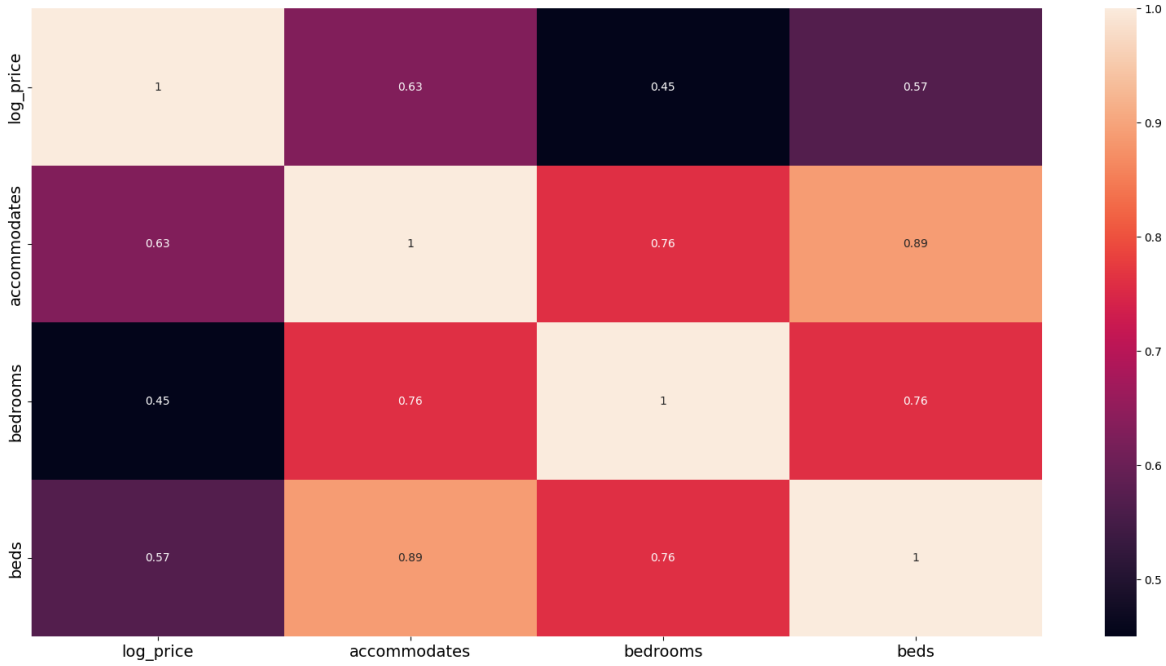


5. Preprocessing

This section rigorously analyzes the Seattle Airbnb dataset, focusing on statistical properties, variable correlations, and relationships between dependent and independent variables. Additionally, the code snippet removes rows with a price of zero for data integrity. The insights gained will inform subsequent analyses for meaningful conclusions specific to the Seattle, Washington Airbnb dataset.

In examining the relationship between log(Price) and various features, the focus shifts to quantitative variables grouped by property characteristics. Numerical features such as accommodates, bedrooms, and beds are considered in this analysis. The correlation matrix,

visualized using a heatmap, provides insights into the degree of correlation between log(Price) and these numerical features. The resulting visualization aids in understanding how property characteristics contribute to the pricing dynamics in the dataset.



In analyzing categorical features such as property_type, room_type, and bed_type, their non-normal distribution in log-transformed price values prompts the application of statistical tests for assessing their significance in predicting prices. The Mann Whitney U (MWU) test is employed for binary response categories, while the Kruskal-Wallis (KW) test is utilized for features with multi-category responses. Both tests aim to determine whether samples from different categories are drawn from the same population, indicating their potential utility in predicting prices.

It is crucial to adhere to minimum sample size criteria for each test. The Mann Whitney U test requires a sample size greater than 20 for each category, and the Kruskal-Wallis test necessitates a sample size exceeding 5 for each category. Prior to analysis, feature engineering steps are implemented for the property_type feature. This includes reassigning instances labeled as 'Bungalow' to the 'House' category and consolidating all other categories under the 'Other' category, focusing on House, Apartment, and Condominium.

The property_type feature is enhanced through feature engineering, reassigning specific instances and creating a simplified version ('property_type_simple'). The simplified version focuses on major categories, and a function, pop_eval, is implemented to conduct statistical tests on log price distributions for features with counts above a specified threshold. The results reveal significant evidence of differences in distributions for both property_type_simple and room_type.

The categorical features, such as instant_bookable, host_response_time, host_is_superhost, host_has_profile_pic, and host_identity_verified, undergo statistical testing using the pop_eval function. Results indicate significant differences in log price distributions for instant_bookable, host_response_time, host_is_superhost, and host_identity_verified. The host_has_profile_pic test, however, does not show significant evidence of different distributions.

The dataset analysis revealed key insights into the distribution and predictors of prices in the context of Airbnb listings. The initial examination of price distribution identified a significant right-skewness, driven by outliers with exceptionally high prices. Applying a log transformation successfully normalized the price distribution. Subsequent correlation analysis identified numerical features such as accommodates, bedrooms, and beds that exhibit a meaningful connection with log-transformed prices.

Moving beyond numerical features, various categorical features emerged as potential predictors for log price. These include property type, room type, bed type, city, instant bookability, host response time, and host identity verification. Additionally, the impact of specific amenities on log price was investigated, revealing key contributors such as free parking, fire extinguisher, microwave, private entrance, and others.

The detailed examination of amenities provides nuanced insights into the factors that significantly influence pricing. These findings can serve as valuable inputs for further analysis or modeling efforts, guiding a more comprehensive understanding of the pricing dynamics within the Airbnb dataset.

6. Machine Learning Modeling

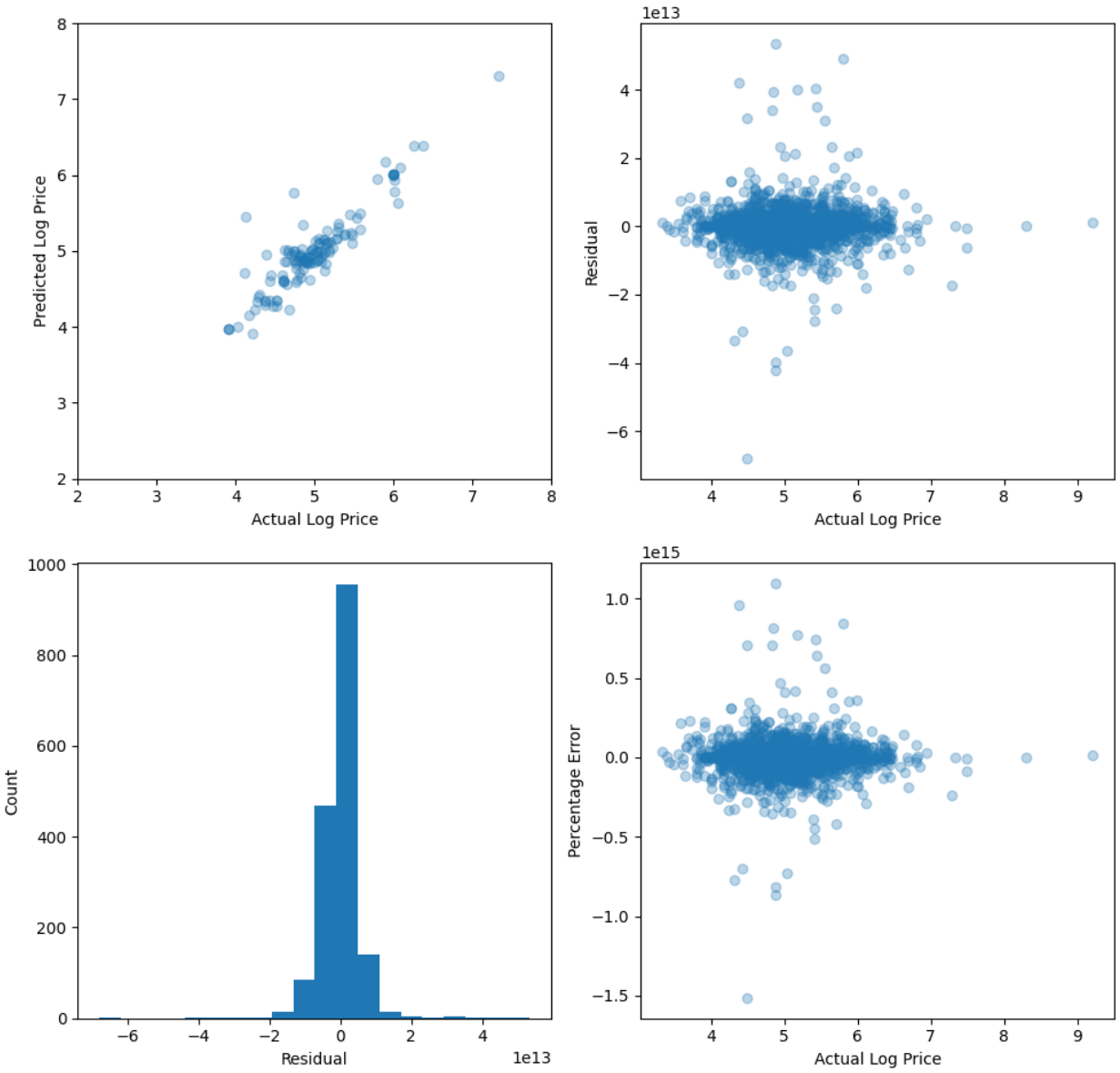
In this section, our focus shifts to developing and evaluating machine learning models for predicting nightly prices. We consider models such as linear regression, random forest, gradient boosting, and extreme gradient boosting (XGB), with log price as the target variable. This choice ensures equal impact on performance metrics for predicting expensive and cheap listings. The log transformation also guarantees non-negative predicted prices, vital for linear regression models. Model selection involves aligning algorithms with dataset characteristics, while hyperparameter

tuning, utilizing techniques like grid search or randomized search, is crucial for optimizing model performance. Our journey involves an iterative process of training, tuning, and evaluation to guide us towards robust solutions that effectively capture the intricacies of our dataset.

In the machine learning model development, the dataset undergoes normalization and is split into training and testing sets using `train_test_split`. The features, excluding `log_price` and `price`, are separated into predictor variables (`X`) and the target variable (`y`). Standardization is applied using `StandardScaler`. The training set is standardized, and the testing set is transformed based on the training set's scaling parameters. This prepares the data for training and evaluating a linear regression model, with metrics like mean squared error (`MSE`) and R-squared (`R2`) used for assessment.

6.1. **Linear Regression Model**

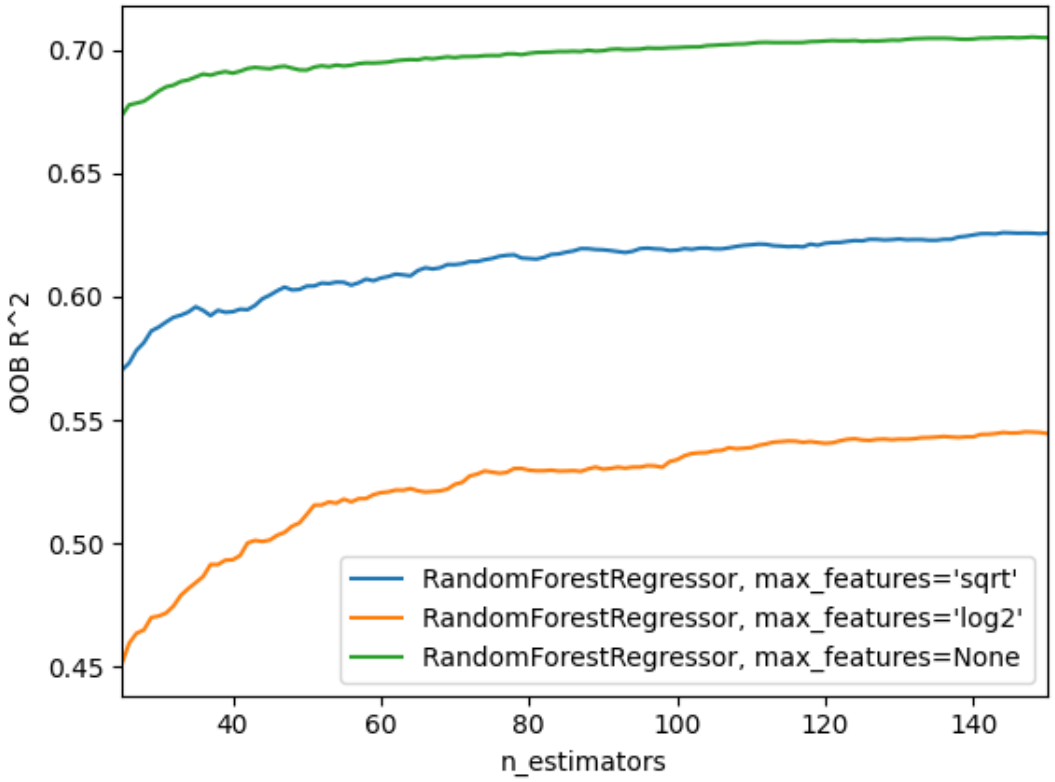
Implemented a linear regression model using scikit-learn's `LinearRegression`. The model performed well on the training data (Train R2: 0.99, Train RMSE: 0.07, Train AAPE: 0.78). However, on the test data, unexpected behavior was observed (Test R2: -1.05e+26, Test RMSE: 6.12e12, Test AAPE: 6.94e13), suggesting potential overfitting or generalization issues. Histograms comparing the distribution of actual and predicted values highlight disparities. Further investigation or model refinement may be needed.



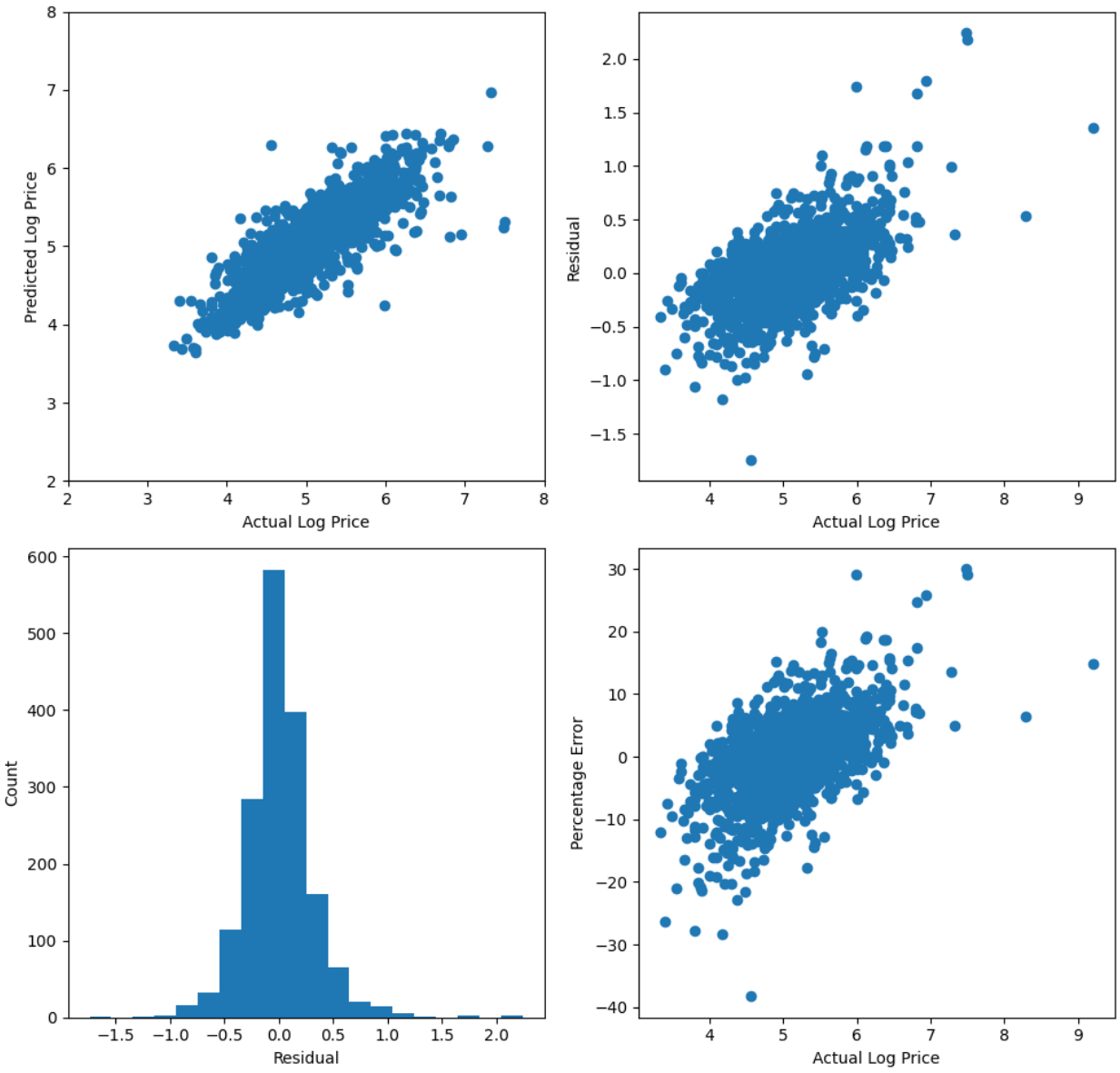
Created diagnostic plots for evaluating the linear regression model on test data. These plots include a scatter plot of predicted vs. actual log prices, a scatter plot of residuals, a histogram of residuals, and a scatter plot of percentage errors. These visuals offer insights into the model's performance and aid in understanding its predictive capabilities.

6.2. **Random Forest**

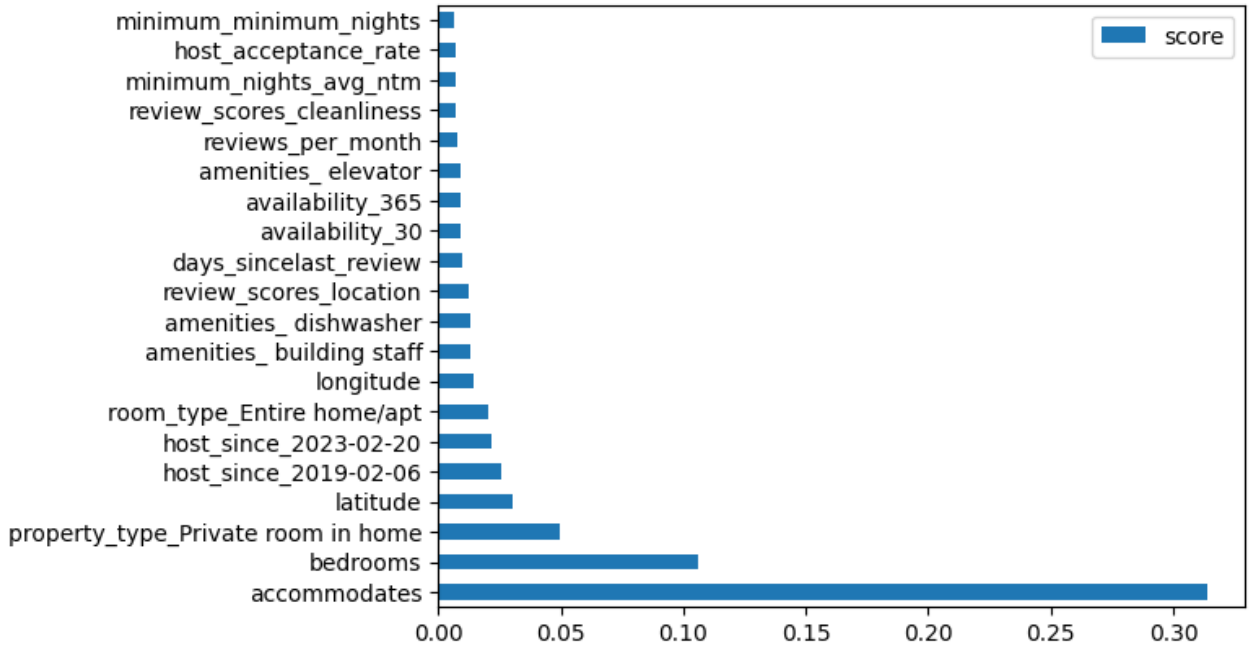
Performed a comprehensive analysis of Random Forest models with varying configurations, including max_features='sqrt', max_features='log2', and max_features=None. Utilized the out-of-bag (OOB) score as a metric to assess model performance across different numbers of estimators. The resulting plot illustrates the OOB R^2 values for each configuration, providing valuable insights into the optimal number of estimators and the impact of max_features on model performance.



The Random Forest regression model is trained and tested, with evaluation metrics such as Test R^2, Test RMSE, and Test AAPE calculated for both log-transformed and unlogged predictions. Visualizations, including scatter plots and histograms, offer a concise overview of the model's performance on the test data.



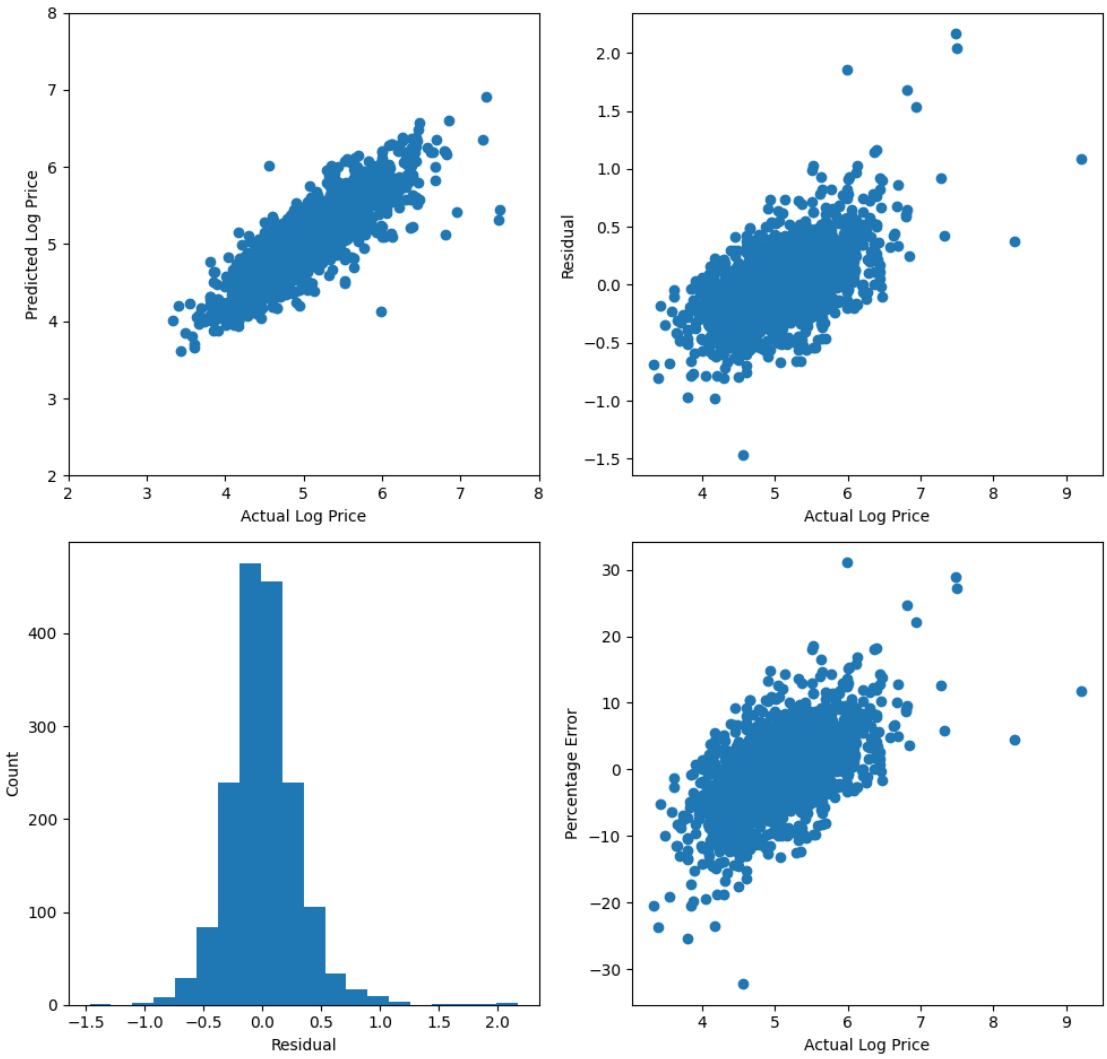
A bar chart showcasing the top 25 feature importance of the Gradient Boosting model is generated. The chart provides insights into the most influential variables influencing the model's predictions, ranked based on their respective importance scores.

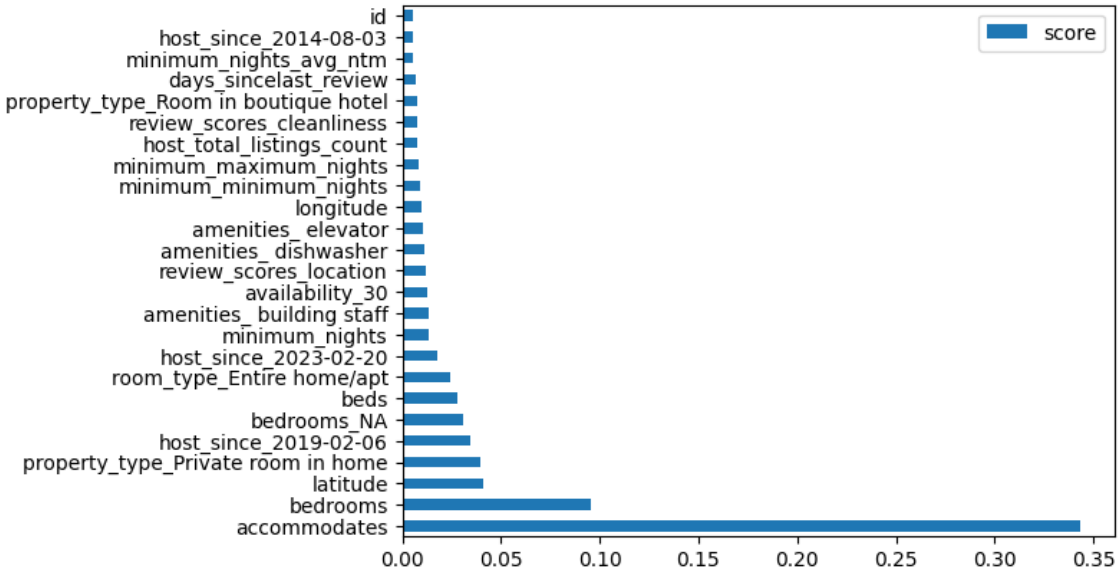


6.3. Gradient Boosting

Gradient boosting, a powerful ensemble learning technique, was employed to predict nightly prices for Airbnb listings in Seattle. Optimal hyperparameters, obtained through grid search with cross-validation, resulted in a Gradient Boosting Regressor with a test R^2 of 0.7487, RMSE of 0.3003, and AAPE of 4.2870. This outperformed the random forest model, showcasing improved explanatory power and reduced prediction errors. Evaluation on unlogged prices yielded a slightly lower R^2 of 0.5847, RMSE of 188.0251, and AAPE of 21.7674, indicating reasonable accuracy. Visualizations highlighted the model's alignment with actual prices, distribution of residuals, and percentage errors. The Gradient Boosting model demonstrates effectiveness in predicting Airbnb prices in Seattle.

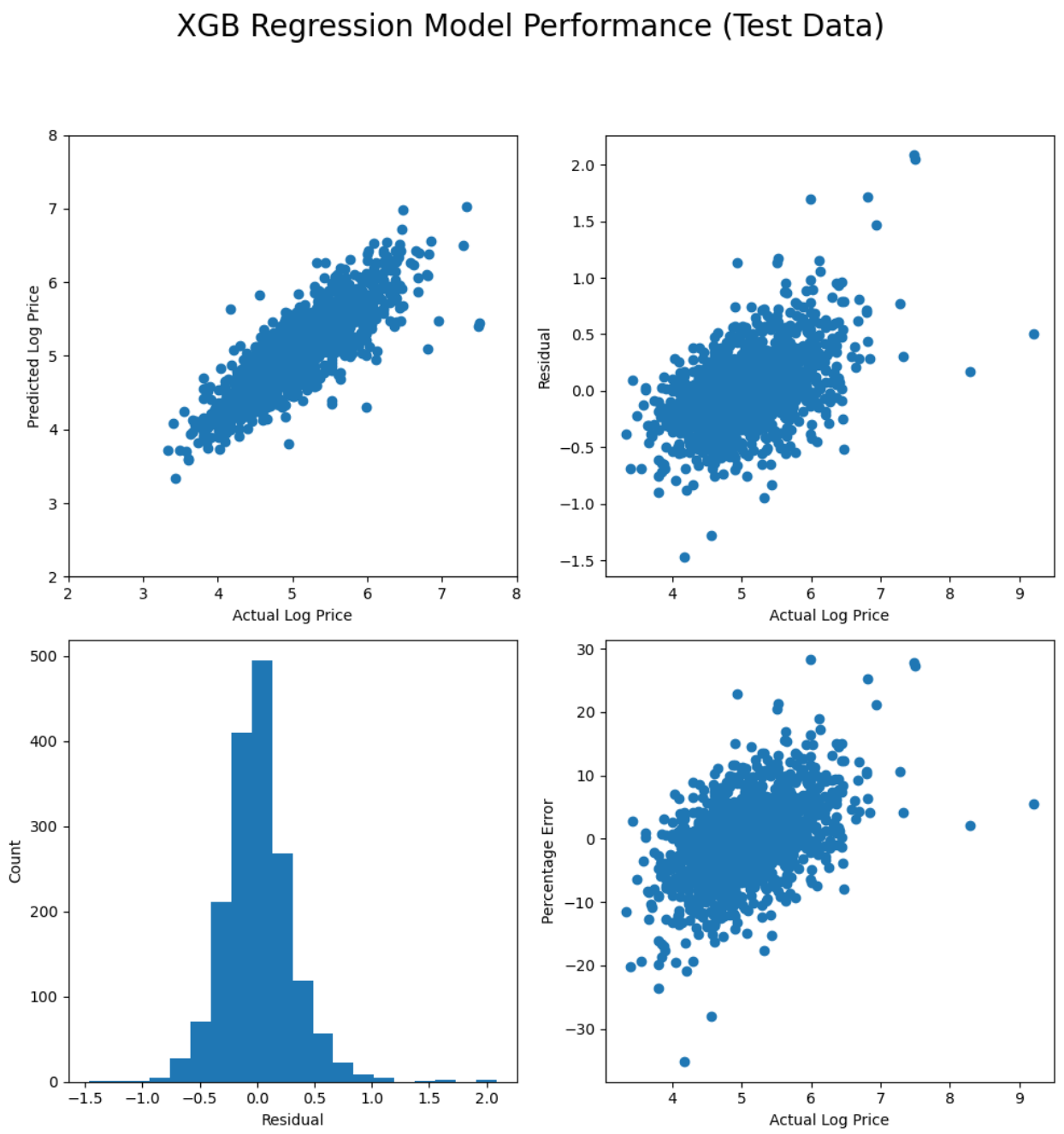
Gradient Boosting Model Performance (Test Data)



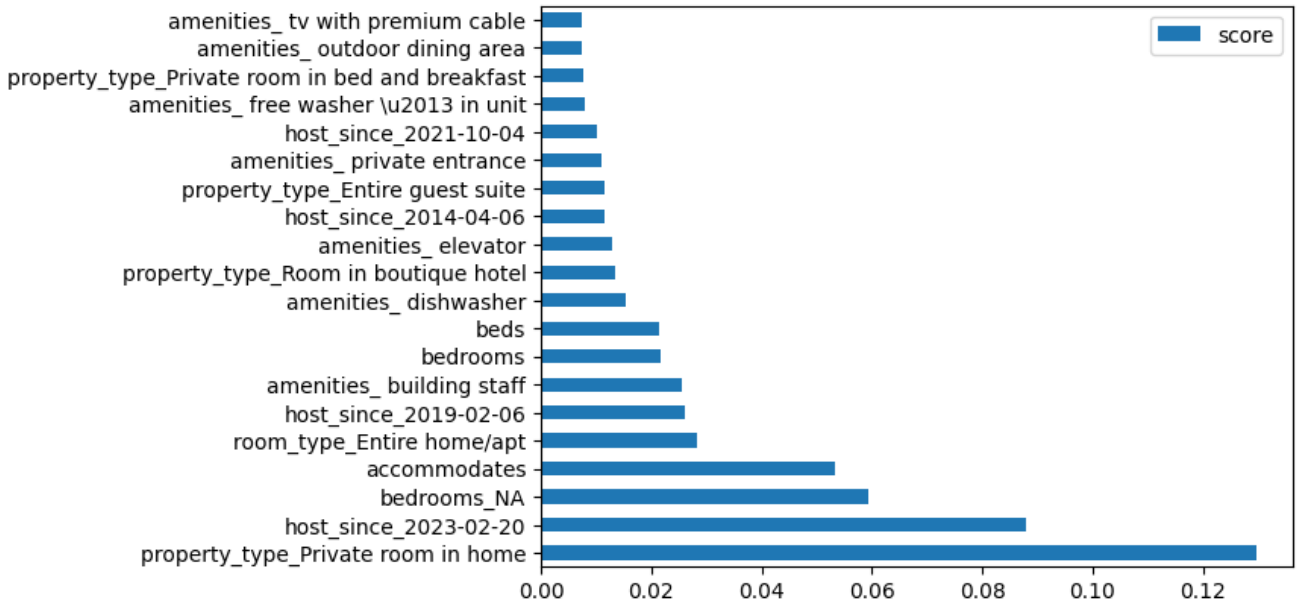


6.4. **XGBoost**

The XGBoost model is tuned using GridSearchCV for parameters like 'colsample_bytree', 'n_estimators', and 'max_depth', resulting in the best parameters: {'colsample_bytree': 1, 'max_depth': 5, 'n_estimators': 100}. The trained model is then evaluated on the test data, yielding a Test R^2 of 0.7429, Test RMSE of 0.3037, and Test AAPE of 4.3577 for log prices. When considering unlogged prices, the model achieves a Test R^2 of 0.7869, Test RMSE of 134.6942, and Test AAPE of 22.2133. Diagnostic plots illustrate the model's performance with scatter plots, residual plots, a histogram of residuals, and a scatter plot of percentage errors.

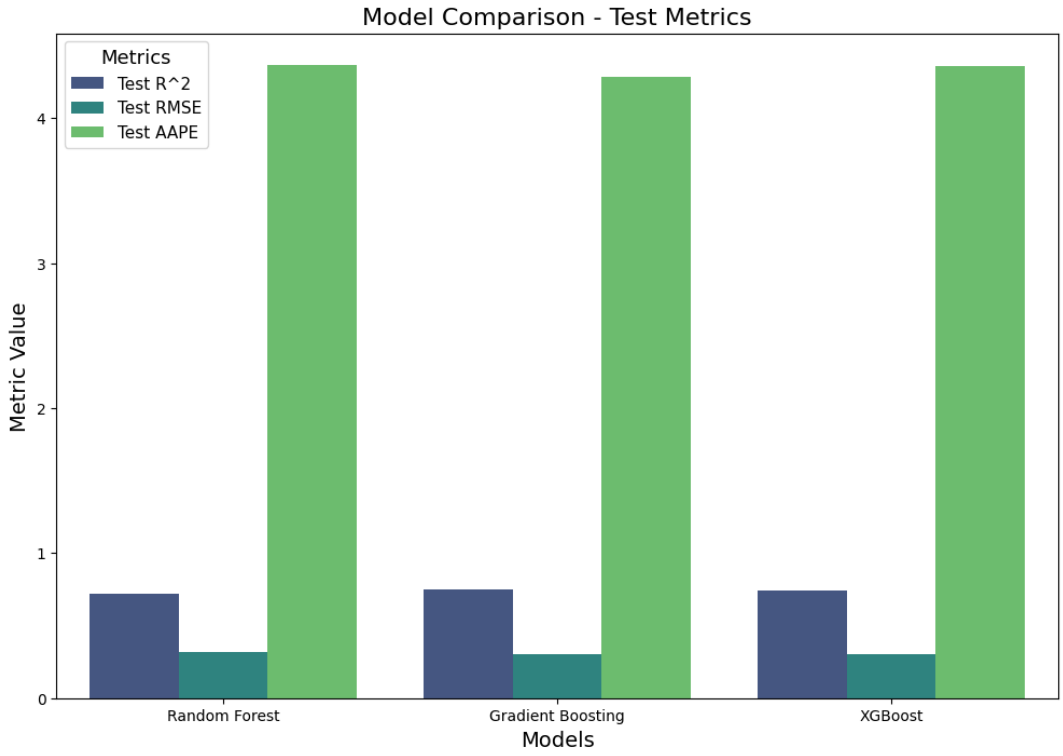


A concise description: Visualizing the top 20 feature importances of the XGBoost model through a horizontal bar chart, excluding 'price' and 'log_price'. The plot illustrates the relative significance of features in influencing the model's predictions.



6.5. Model Comparison

The model comparison chart provides a concise overview of the performance metrics for three different regression models—Linear Regression, Random Forest, Gradient Boosting, and XGBoost. The metrics include Test R^2, Test RMSE, and Test AAPE. Notably, the Linear Regression model exhibits extreme values in the Test R^2 and Test RMSE metrics, indicating potential issues with its performance. On the other hand, Random Forest, Gradient Boosting, and XGBoost models show competitive and reasonable performance across all three metrics, with Test R^2 values ranging from 0.72 to 0.75, Test RMSE values around 0.30, and Test AAPE values approximately between 4.29 and 4.37. The visualization aids in quickly comparing the models and identifying their strengths and weaknesses in predicting the target variable.



7. CONCLUSION

In our pursuit of predicting nightly prices for Airbnb listings in Seattle, Washington, we executed a meticulous data science workflow encompassing Exploratory Data Analysis (EDA), data preprocessing, feature engineering, and evaluation of diverse machine learning models. During EDA, we gained insights into feature distributions and relationships, followed by addressing missing values, outliers, and crucial data transformations in the preprocessing phase. Feature engineering enhanced model predictive capabilities by extracting meaningful features and introducing new variables. We explored linear regression, random forests, gradient boosting, and XGBoost, considering their potential to capture intricate patterns.

Hyperparameter tuning using grid search focused on key parameters such as learning rate, max tree depth, and the number of features. Model comparison results revealed the performance metrics for Random Forest, Gradient Boosting, and XGBoost. A grouped bar chart visually illustrated the comparative model performance, emphasizing Gradient Boosting's highest R^2 value as the superior model in explaining nightly price variability.

Gradient Boosting emerged as the best model, not only due to its leading R^2 but also competitive RMSE and AAPE metrics. The ensemble learning nature of Gradient Boosting, combining multiple weak learners, contributed to its robust performance. In conclusion, our comprehensive

data science approach and detailed model comparison recommend Gradient Boosting as the preferred model for predicting nightly prices in the specific Airbnb dataset.

REFERENCES:

[1] Rezazadeh Kalehbasti, P., Nikolenko, L., & Rezaei, H. (2021, August). Airbnb price prediction using machine learning and sentiment analysis. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction* (pp. 173-184). Cham: Springer International Publishing.

[2] Gangarapu, S., & Mernedi, V. S. A. (2023). Predicting Airbnb Prices in European Cities Using Machine Learning.

[3] Luo, Y., Zhou, X., & Zhou, Y. (2019). Predicting airbnb listing price across different cities.

[4] Yang, S. (2021, March). Learning-based airbnb price prediction model. In *2021 2nd International Conference on E-Commerce and Internet Technology (ECIT)* (pp. 283-288). IEEE.

[5] Choudhary, P., Jain, A., & Baijal, R. (2018). Unravelling airbnb predicting price for new listing. *arXiv preprint arXiv:1805.12101*.

[6] Kirkos, E. (2022). Airbnb listings' performance: Determinants and predictive models. *European Journal of Tourism Research*, *30*, 3012-3012.

[7] Tang, E., & Sangani, K. (2015). Neighborhood and price prediction for San Francisco Airbnb listings. *Departments of Computer science, Psychology, economics–Stanford University*.

[8] Haldar, M., Abdool, M., Ramanathan, P., Xu, T., Yang, S., Duan, H., ... & Legrand, T. (2019, July). Applying deep learning to airbnb search. In *proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & Data Mining* (pp. 1927-1935).

[9] Jain, S., Proserpio, D., Quattrone, G., & Quercia, D. (2021). Nowcasting gentrification using Airbnb data. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1-21.

[10] Afrianto, M. A., & Wasesa, M. (2020). Booking prediction models for peer-to-peer accommodation listings using logistics regression, decision tree, K-nearest neighbor, and random Forest classifiers. *Journal of Information Systems Engineering and Business Intelligence*, *6*(2), 123-32.

[11] Garcia, S. G. (2023). *Evaluating the Impact of Image Features on Airbnb Price Predictions: A Machine Learning Approach to Hedonic Pricing* (Master's thesis).