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

Springboard- Data Science Career Track

Pramod Acharya

Springboard

Mentor: Kenneth Gil Pasquel

Problem Identification

Specific Problems with Airbnb Listing Prices:

•**Data Overload:** A 2021 survey (https://www.hostaway.com/) found that **72% of hosts** feel overwhelmed by the data presented in "Smart Pricing," making it difficult to set competitive prices.

•**Limited Control:** A 2022 Airbnb Community Forum thread (https://community.withairbnb.com/t5/Help/Airbnb-Price-Error/m-p/882653) with **over 100 responses** highlights hosts' frustration with the limited control over pricing offered by "Smart Pricing.”

•**Unintended Price Changes:** A 2023 Airbnb Community Center report (https://www.airbnb.com/help/article/2718) details a technical glitch causing **unexpected price drops** for multiple hosts, leading to lost bookings.

•**Hidden Fees:** A 2023 NerdWallet: https://www.nerdwallet.com/ study revealed that **54% of guests** find the total cost on Airbnb unclear due to hidden fees, leading to booking dissatisfaction.

•**Lack of Communication:** According to a 2022 Forbes: https://www.forbes.com/ article, many hosts **criticize the lack of communication** regarding "Smart Pricing" adjustments, making it difficult to understand pricing fluctuations and adjust strategies.

Objective

- Accurately predict dynamic Airbnb prices in Seattle.

Factors

- Custom pricing, availability, demand, and additional charges.

Challenges:

- Dynamic pricing variations from initial listings

1. **Airbnb Evolution:**
    1. Global business since 2007, $2.6B revenue in 2017.
2. **Host Pricing Challenge:**
    1. Balance crucial for customer attraction and revenue maximization.
3. **Project Focus:**
    1. Data analytics and ML for accurate pricing in Seattle.
4. **Factors Impacting Prices:**
    1. Custom pricing, availability, demand, and guest charges.
5. **Goals and Structure:**
    1. Develop robust pricing model for Seattle, enhance user experience.

- **What factors fluctuate the listing price?**
- **Is it possible to accurately predict the listing prices?**

Data Collection

•Acquired from insideairbnb.com, a site scraping Airbnb's database periodically.

•Specific file analyzed: 'listings.csv' downloaded from http://insideairbnb.com/get-the-data/.

•Dataset last scraped on 09/18/2023, ensuring current and representative data for Seattle.
Dataset Overview:

•Represents Airbnb listings in Seattle, Washington, with 6,823 rows and 75 features.
•Each row corresponds to a unique Airbnb listing.

2.1. Property Information:
•Unique identifier
•Name
•Description
•Property type
•Room type
•Accommodates
•Bedrooms
•Beds

Host Information:
•Host ID
•Host name
•Host since
•Host location
•Host about
•Host response time
•Host response rate
•Host acceptance rate
•Host is superhost

Price and Availability:
•Nightly rate
•Minimum and maximum nights
•number of available days in the next 30, 60, 90, 365 days

Review Information:
•Total number of reviews
•Overall rating score

Web Scraping Information:
•Scrape ID
•Last scraped date

Data Wrangling

Data Wrangling Overview:

- Essential phase for effective analysis of Airbnb listing prices in Seattle, focusing on correcting data types, handling missing data, creating new features, and organizing the dataset for meaningful analyses.

Tailored Steps for Seattle Listings:

- Validate and convert 'price' to numeric format, address missing values, create new features like 'price per bedroom,' and handle outliers for a cleaner, more reliable dataset.

Handling Nightly Prices:

- Identified outliers exceeding $5000.
- Recorded details of high-priced listings.
- Created actual_price for recording actual prices.
- Dropped redundant column host_total_listings_count.

Host Information:

- Examined host_listings_count and host_total_listings_count.
- Dropped redundant column if identical.

Distribution of Prices by Room Type:

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

- **Geographical Information:**
    - Investigated latitude and longitude.
- **Property Information:**
    - Explored accommodates, bedrooms, and beds.
    - Identified extreme property values.

- **Datetime Features:**
    - Examined datetime columns like host_since, first_review, and last_review.
    - Imputed missing values in datetime features.
- **Correlation Analysis:**
    - Generated correlation matrix for relevant columns.
    - Visualized correlations with a heatmap.
    - Identified pairs with correlation coefficients > 0.75 and created scatter plots.
- **New Feature Creation:**
    - Created amenities_list from the amenities column.
    - Generated a set of unique amenities for exploration.

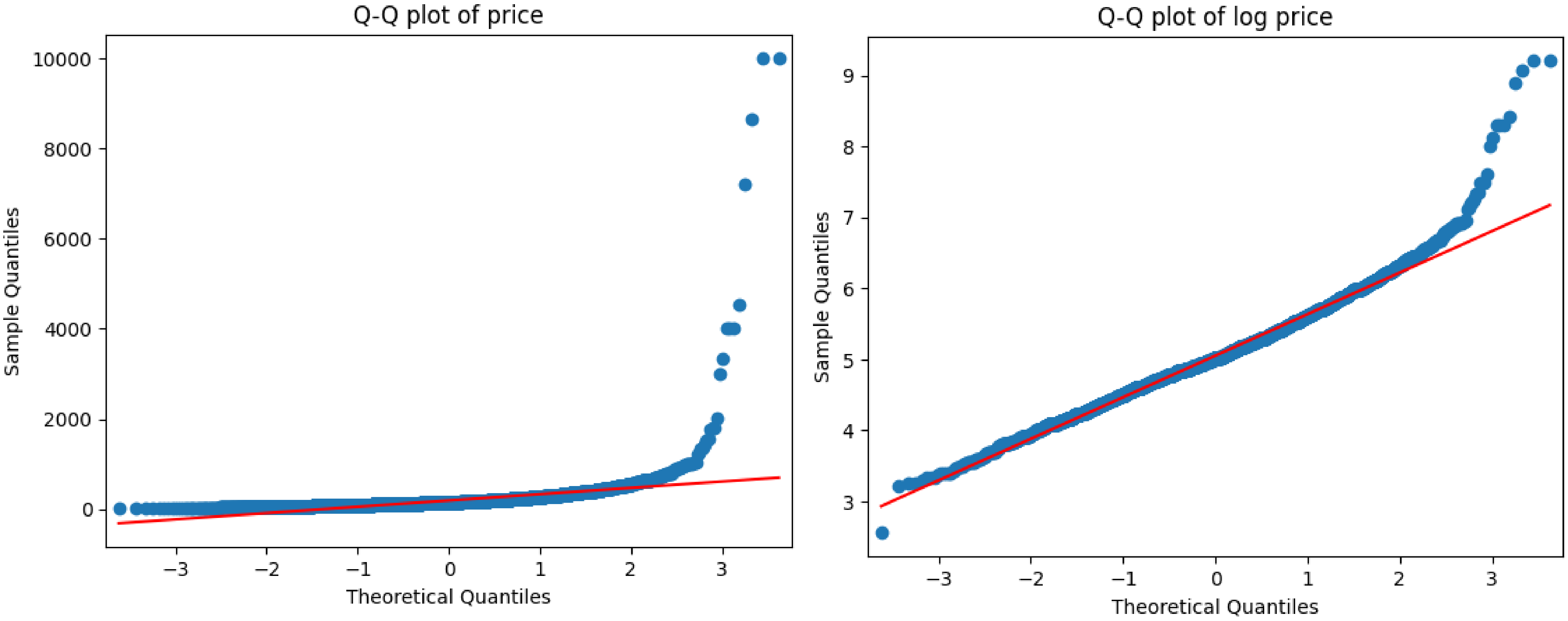
- **Booking Policy:**
    - Investigated minimum nights distribution.
    - Addressed unusual maximum nights value of 999.
- **Availability:**
    - Explored availability metrics.
    - Examined availability_30, availability_60, availability_90, and availability_365.
- **Reviews:**
    - Explored review-related features.
    - Examined number_of_reviews, number_of_reviews_ltm, and various review_scores.
- **Missing Values Imputation:**
    - Examined missingness in each row.
    - Imputed missing numeric features like host_response_rate, host_acceptance_rate, bedrooms, and beds.
    - Imputed missing categorical features like description, neighborhood_overview, host_location, etc.

Exploratory Data Analysis

- **Nightly Price Analysis:**
  - 'clean_df' (6822 entries) shows right-skewed distribution with mean $193.97 and std $275.43.
  - Log transformation applied due to outliers (max $10,000), resulting in average nightly price of $194 (IQR $112).

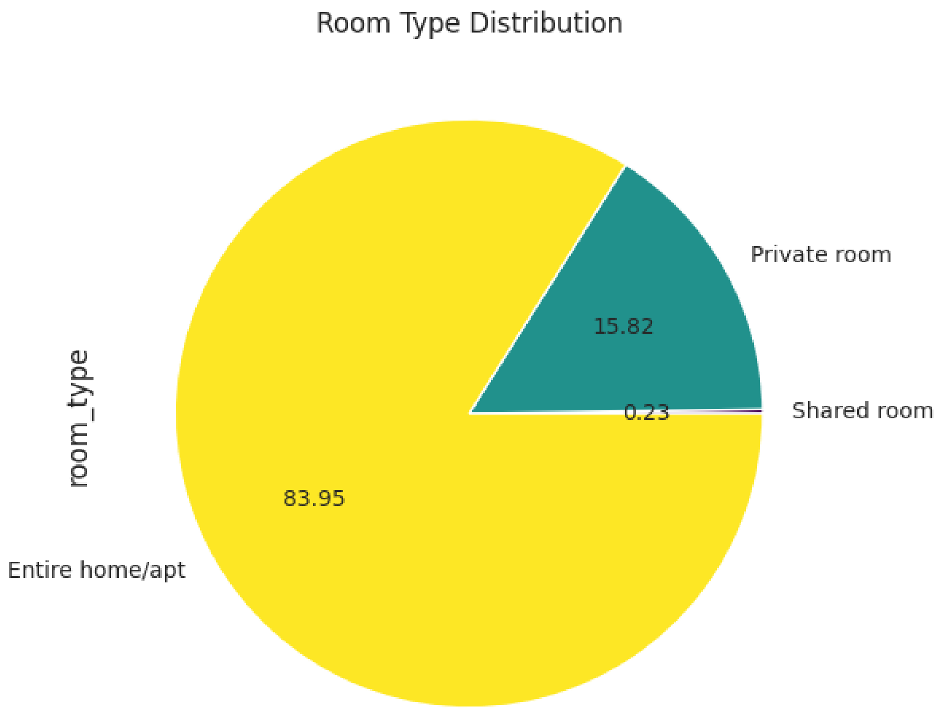
- **Refined Property Types:**
  - Focus on high-count types (threshold: 3%) in 'df_top_property_type.'
  - Seaborn boxplot reveals pricing dynamics and trends, addressing granularity challenges.

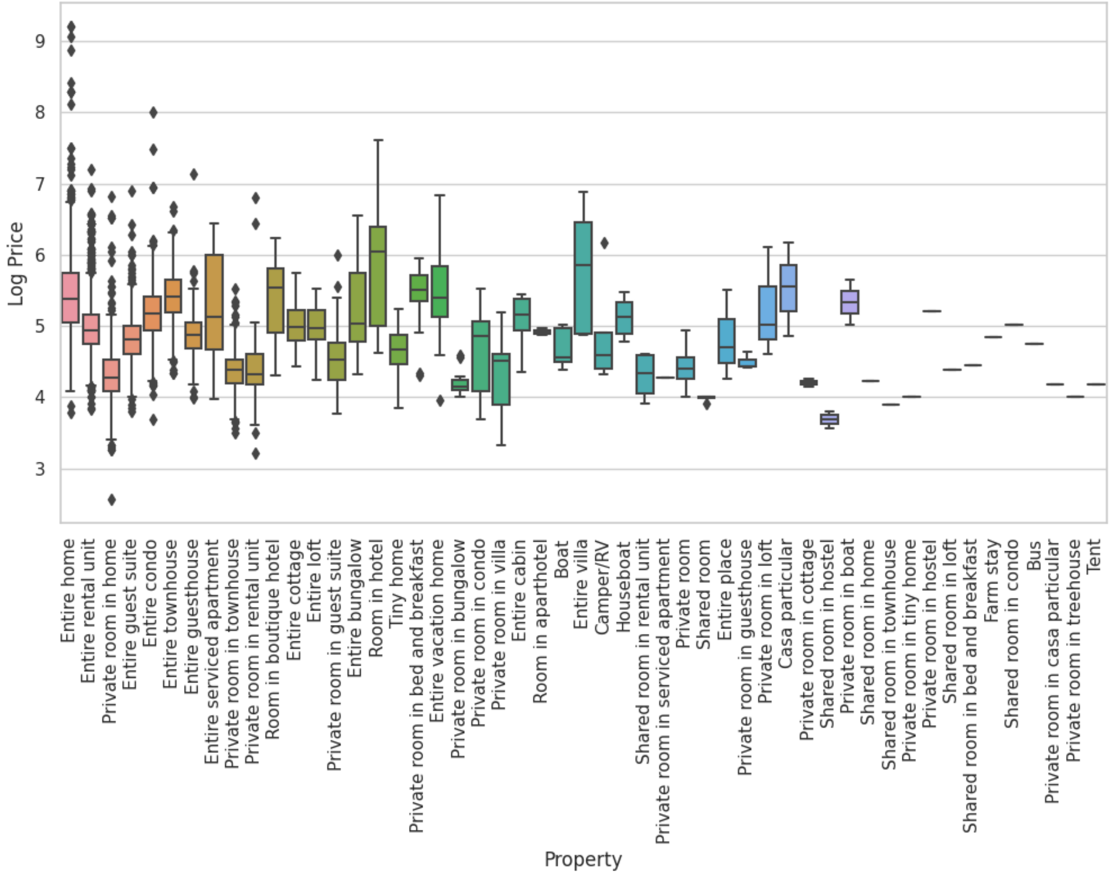
- **Key Findings:**
  - Log transformation enhances analysis accuracy.
  - 'df_top_property_type' captures relevant info: 'property_type,' 'log_price,' and 'room_type.'
  - Visualizations offer insights into specific property types' influence on pricing dynamics.

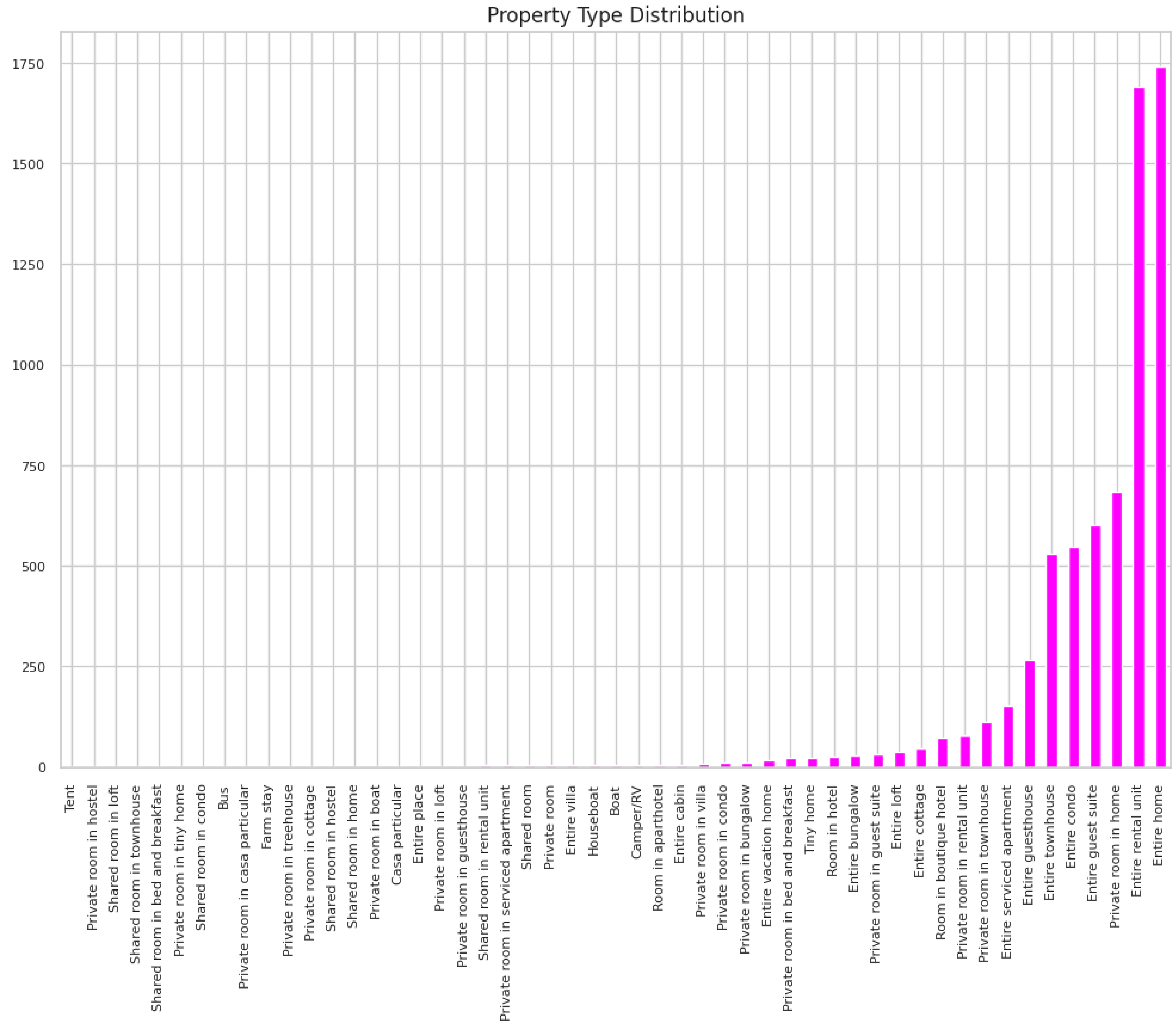

Q-Q plot of price
Sample Quantiles
10000
8000
6000
4000
2000
0
−3
−2
−1
0
1
2
3
Theoretical Quantiles
Q-Q plot of log price
Sample Quantiles
9
8
7
6
5
4
3
−3
−2
−1
0
1
2
3
Theoretical Quantiles

Property Information

- **Refined Property Type Examination:**
  - Focus on property types with substantial listings (threshold: 3%).
  - 'df_top_property_type' dataset captures 'property_type,' 'log_price,' and 'room_type,' visualized with Seaborn boxplot to reveal pricing variations and trends.



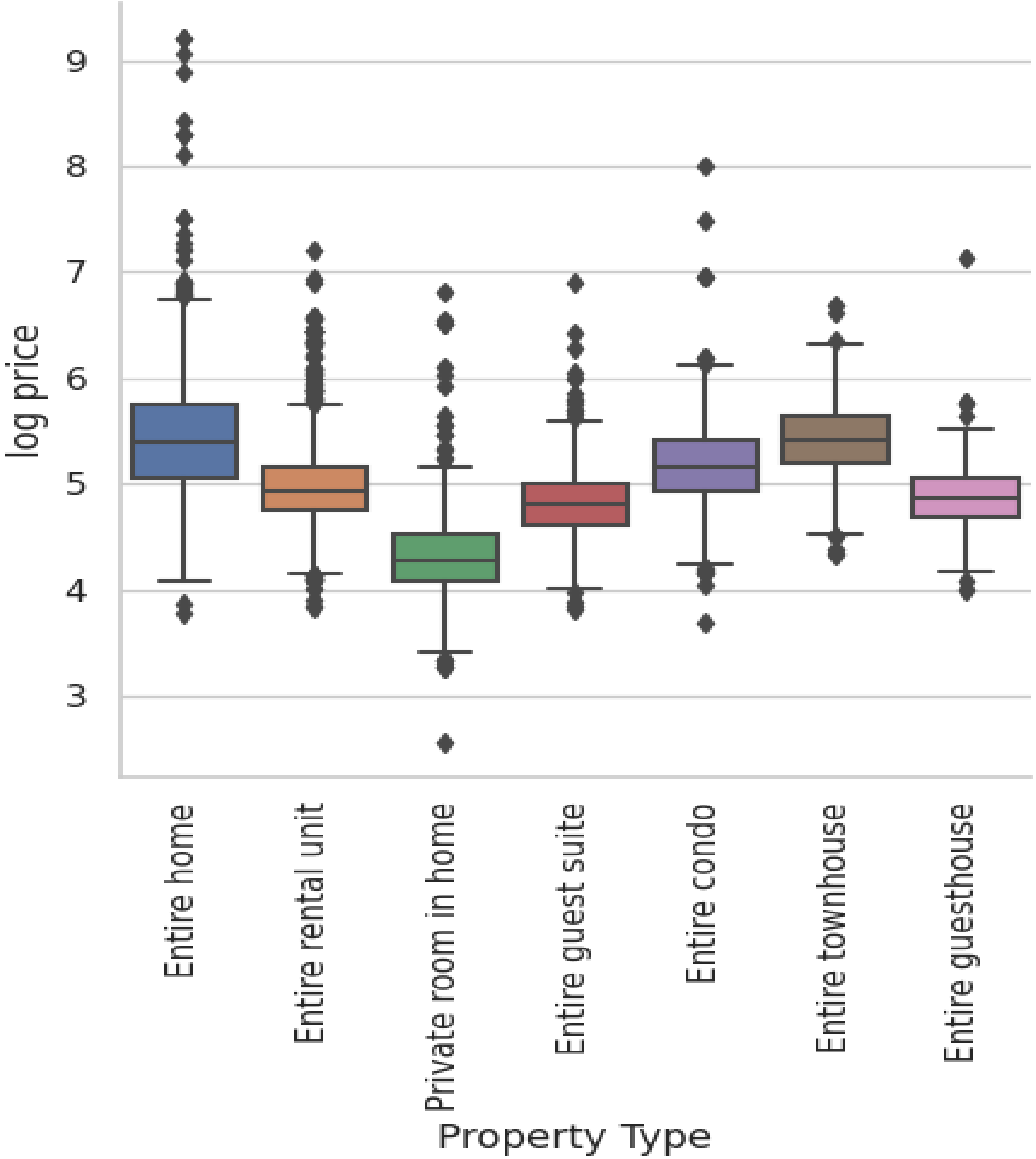


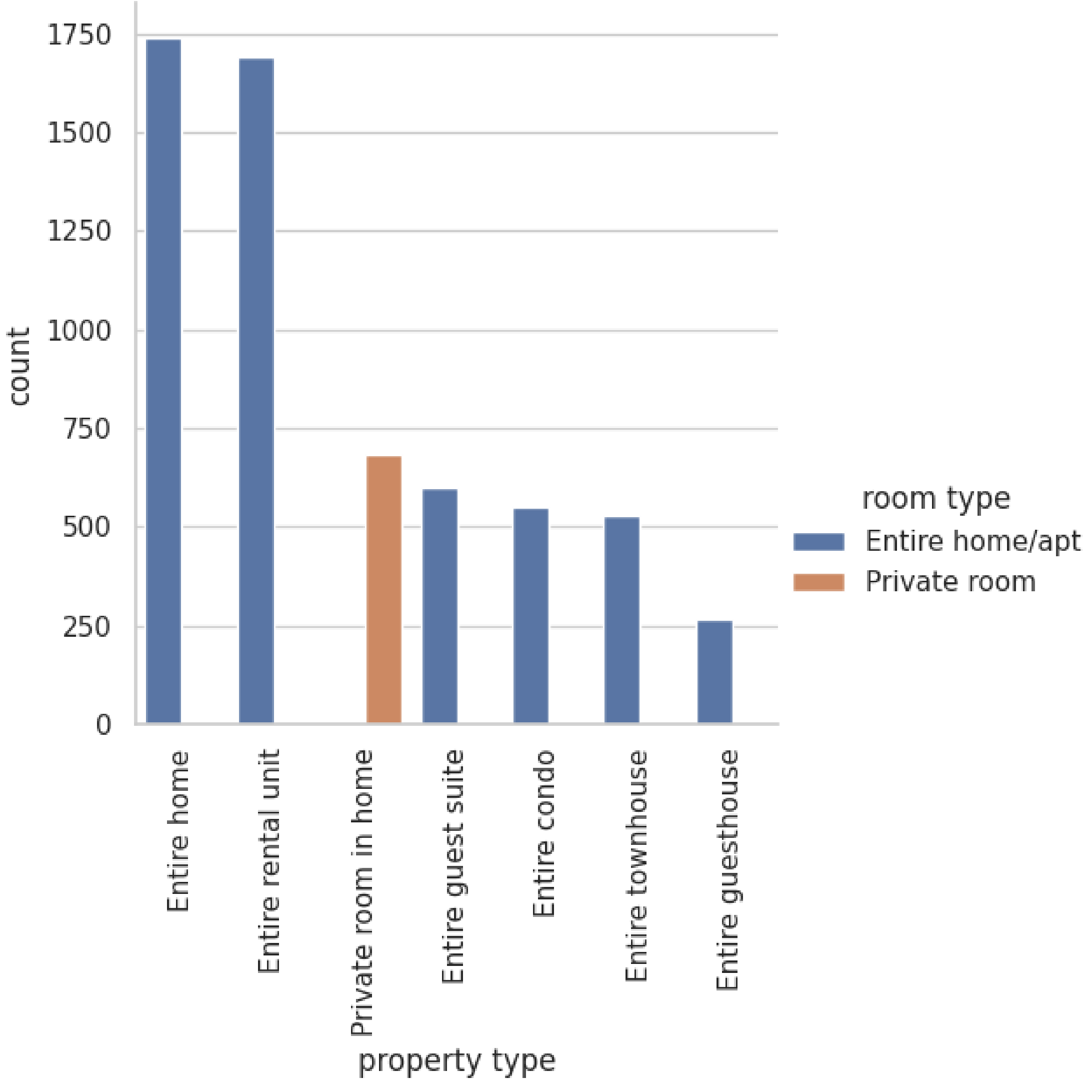


Property Type Insights:

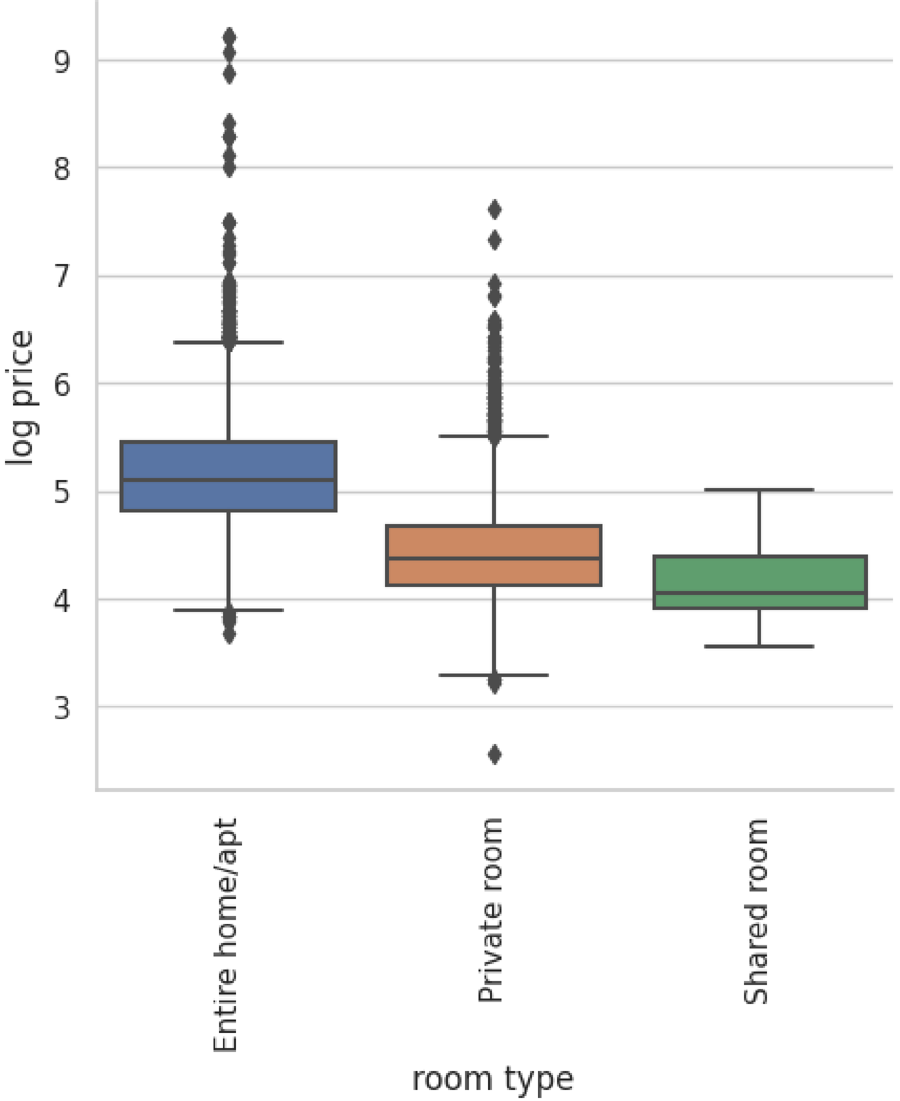
- Approximately 1,700 entire residences, double the count of private rooms (700), indicating dominance of apartments and houses.

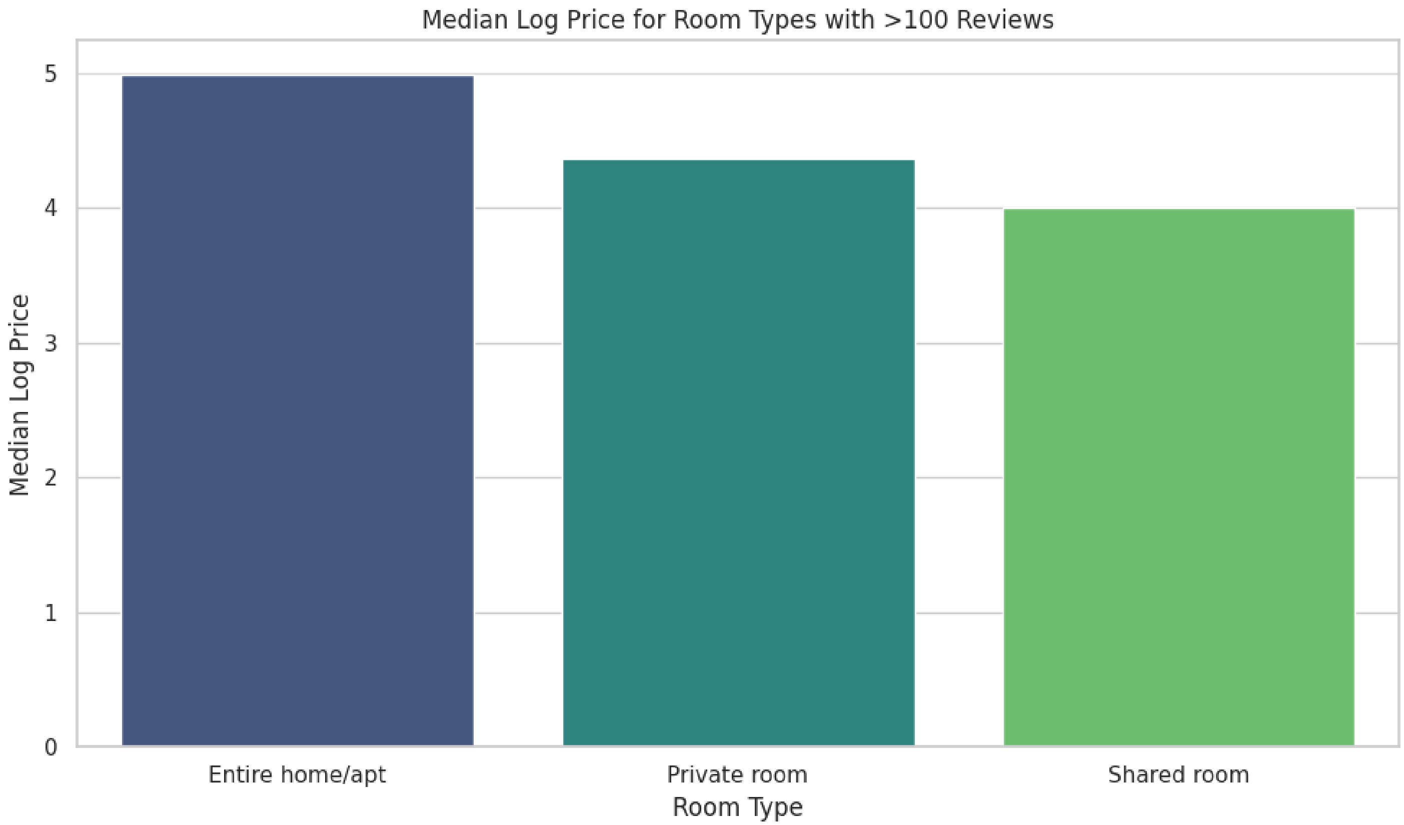
- Emphasizes the pivotal role of property and room types in predicting Airbnb listing prices.


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
Property Type

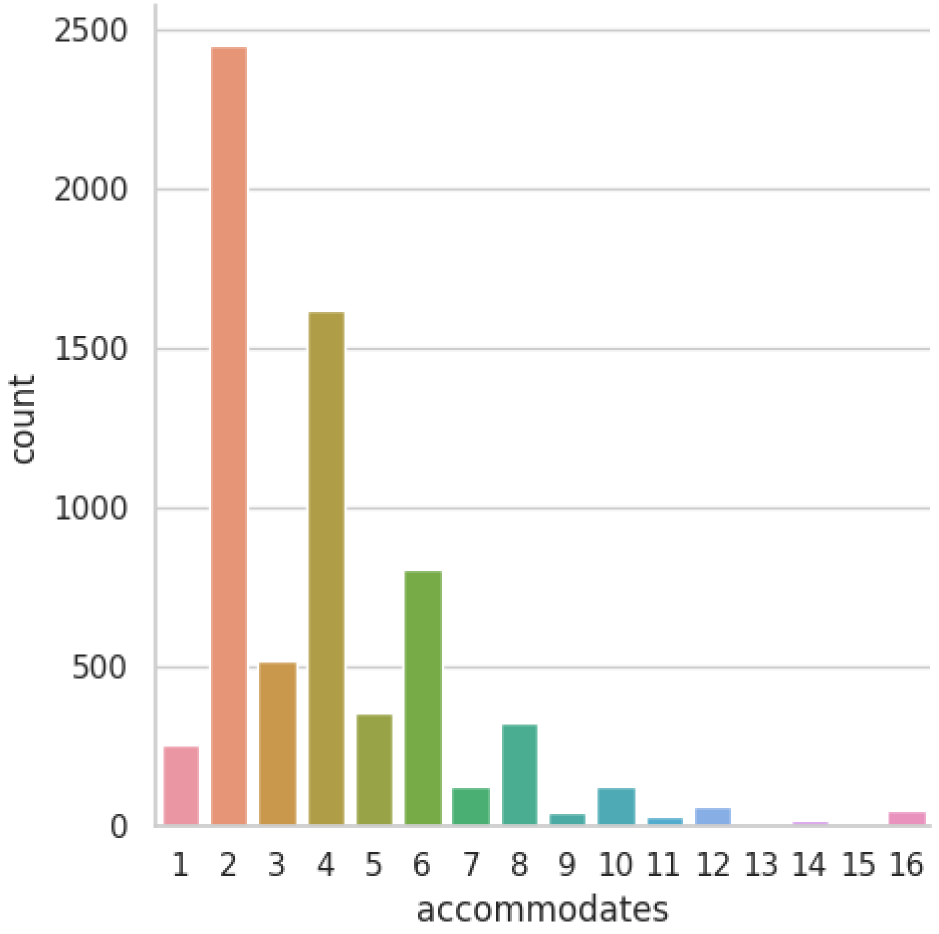

count
1750
1500
1250
1000
750
500
250
0
Entire home
Entire rental unit
Private room in home
Entire guest suite
Entire condo
Entire townhouse
Entire guesthouse
property type
room type
Entire home/apt
Private room

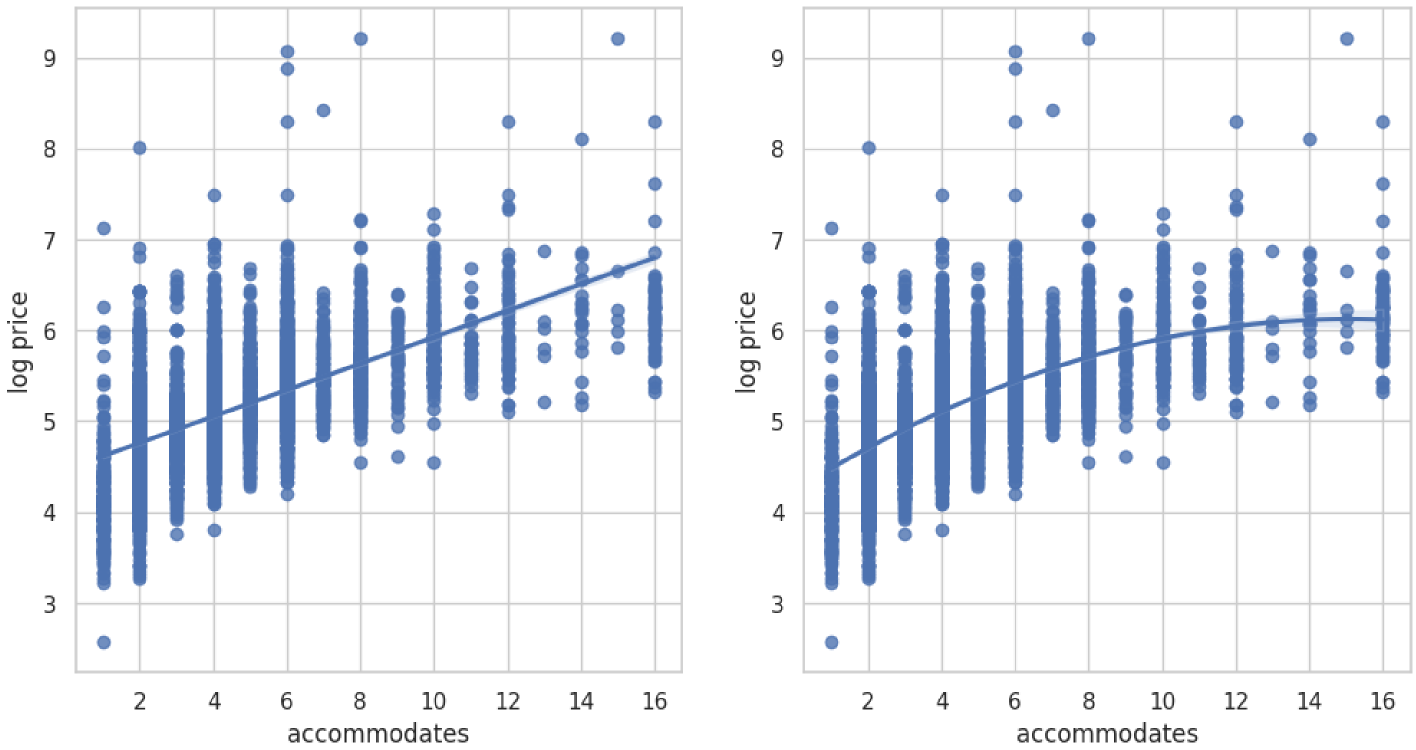
Available Room Types




Median Log Price for Room Types with >100 Reviews

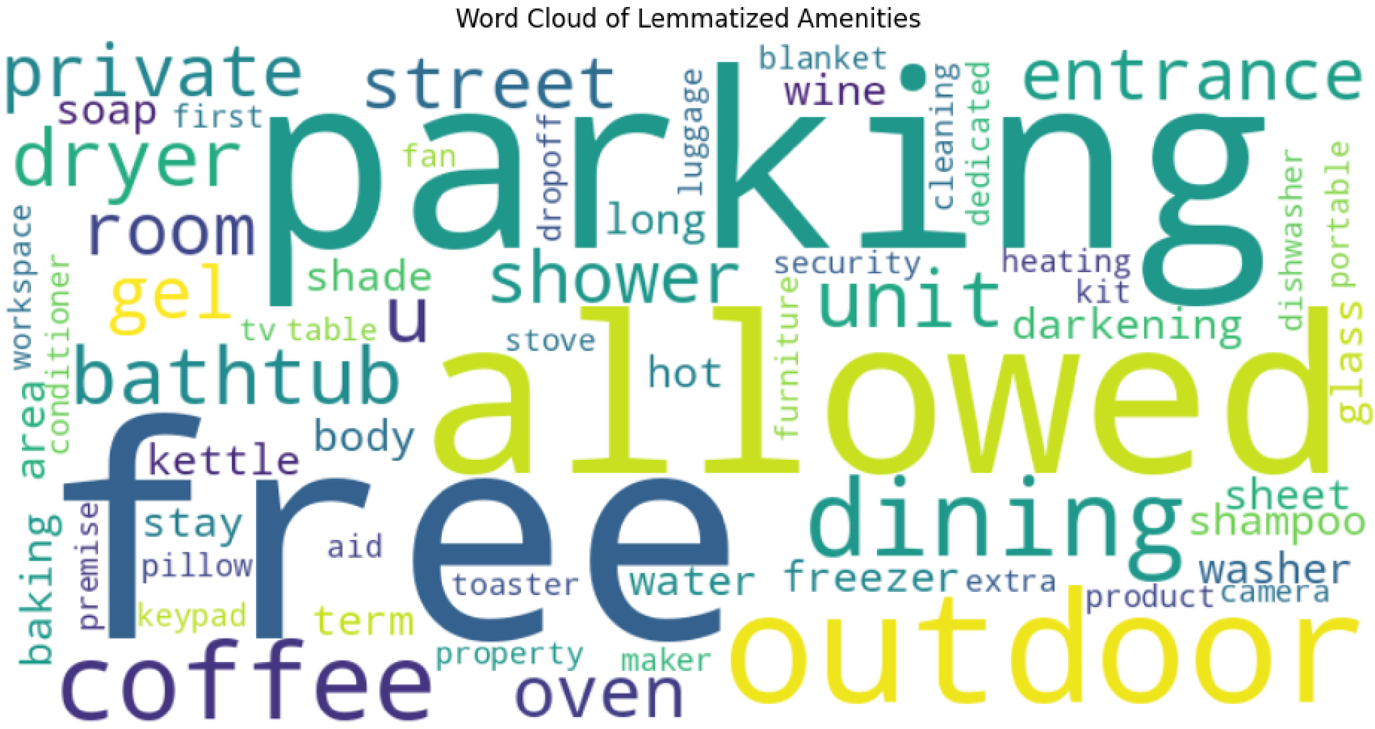
Distribution on Accommodates

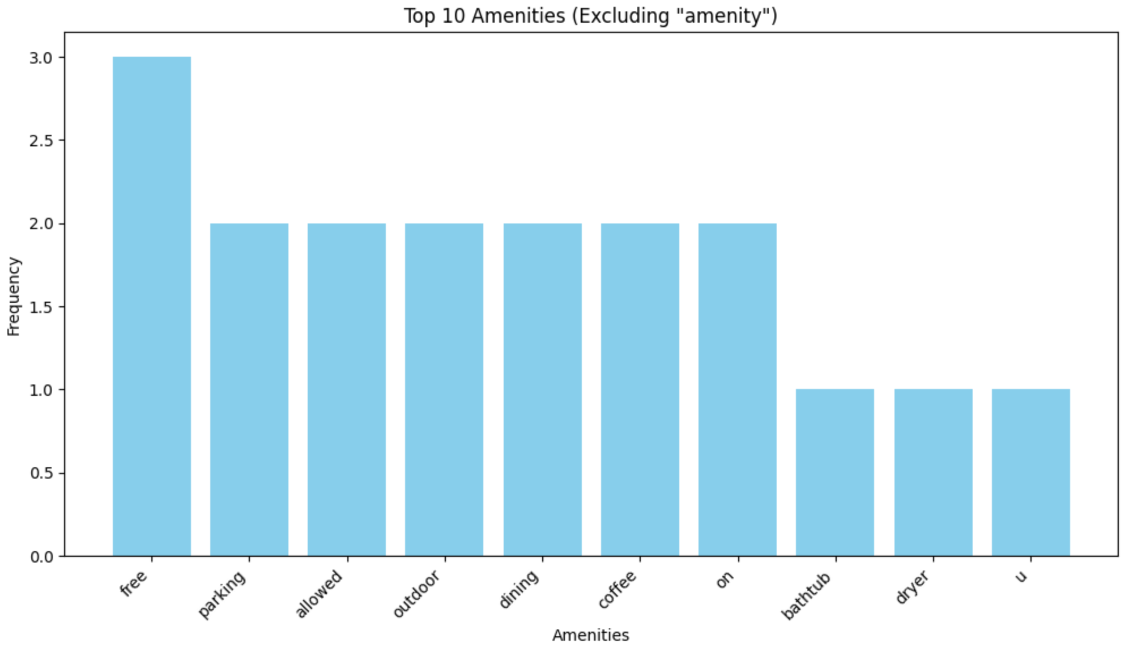




Amenities

- **Amenities Impact on Pricing:**
    - Top 20 common amenities range from essentials to specialized features.
    - Identification of top 10 impactful amenities sheds light on influential features affecting pricing dynamics.
- **Popular Amenities Insights:**
    - Analysis reveals the 20 most common amenities, crucial for understanding pricing impact and consumer preferences.
    - Code generates a word cloud and bar chart for a quick visual overview of prevalent amenities in listings.

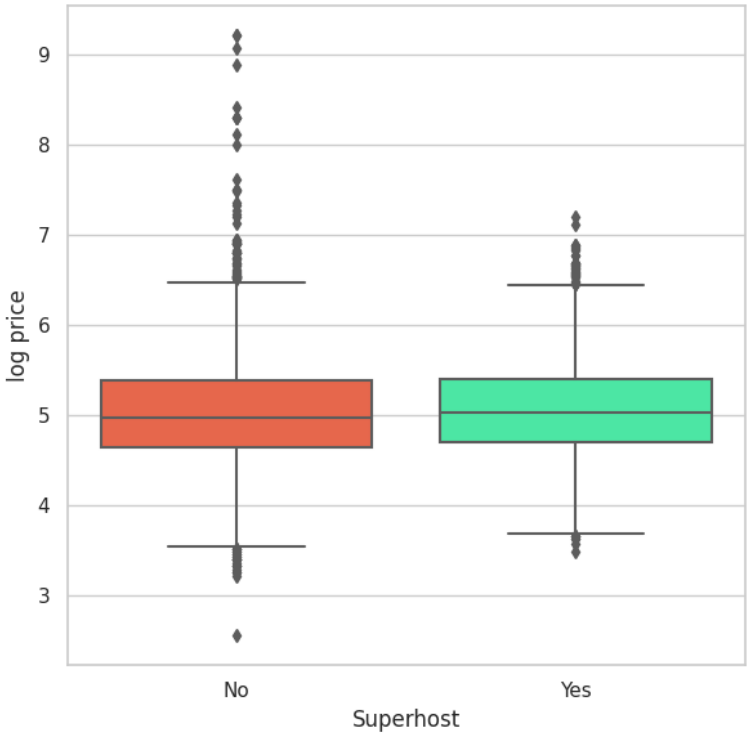


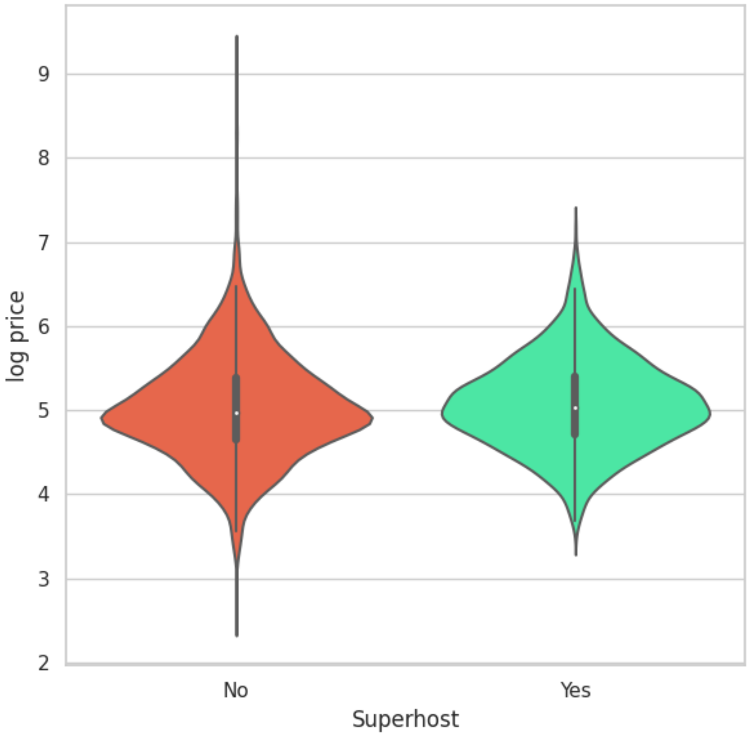


Host Information

•46% of listings are from superhosts, known for exceptional service.

•Boxplots and violinplots compare log prices, suggesting similar median prices with some upper quartile variations, indicating potential pricing differences between superhosts and non-superhosts.





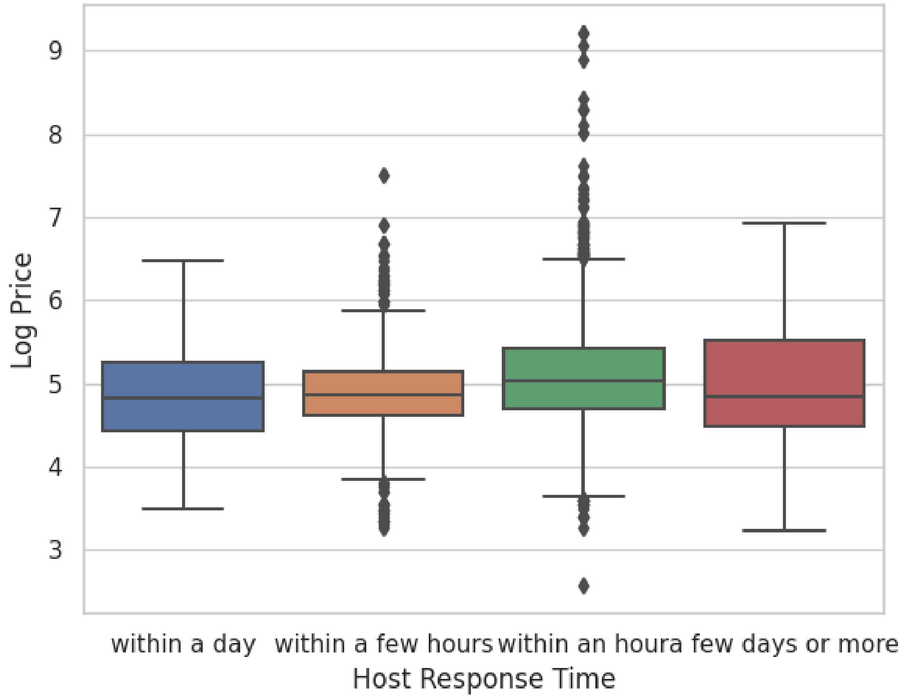
•Examines host longevity, review count, and log prices for data points with over 300 reviews.

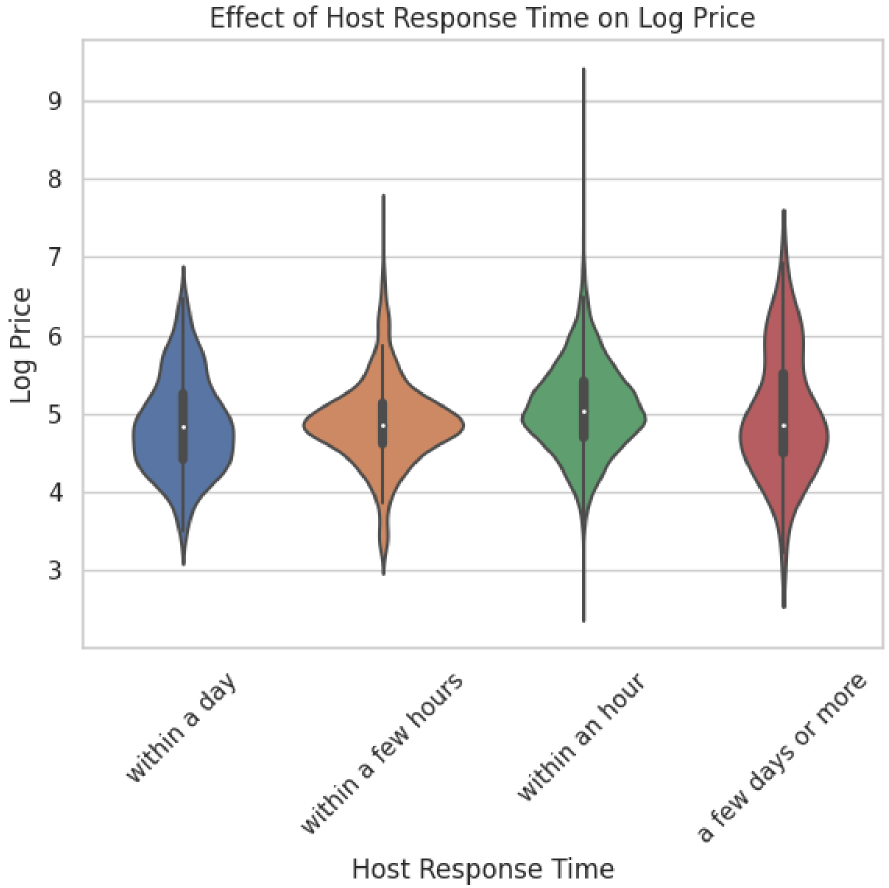
•Visual representation offers insights into the correlation between these factors and pricing dynamics.

- Nearly 85% of listings have hosts responding within an hour, highlighting their commitment to prompt communication.

- Boxplots and violinplots show consistent median log prices across various response times,

- It offer insights into the potential relationship between host response time and listing prices.

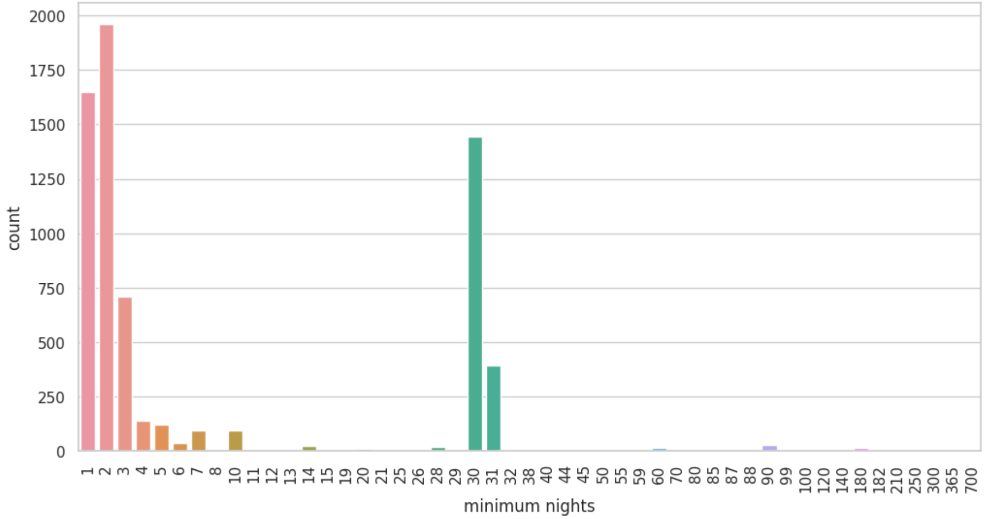


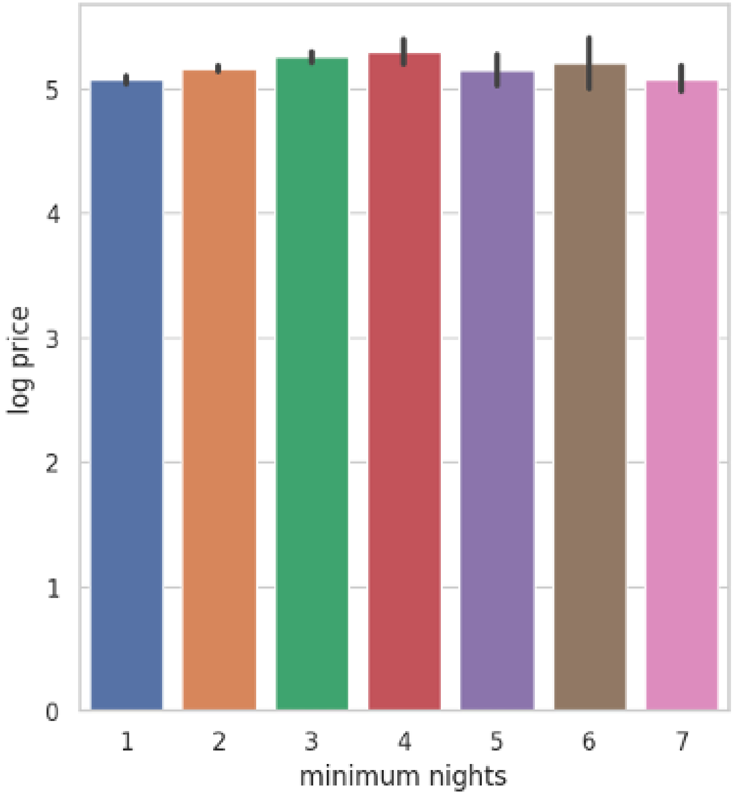

Effect of Host Response Time on Log Price

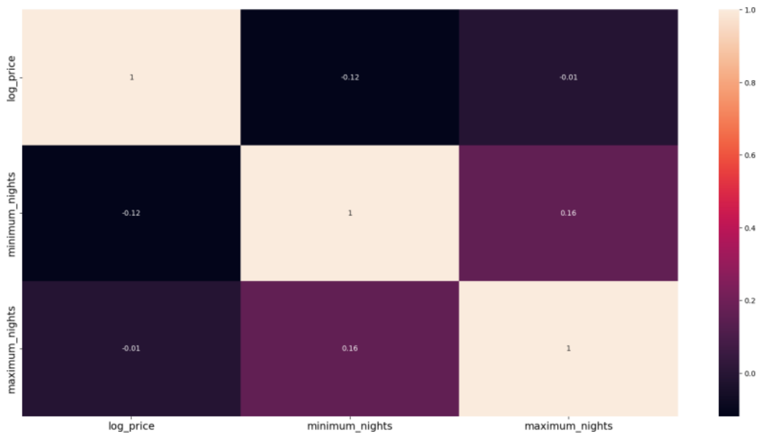
Minimum Night Stay

- Mean minimum nights: around 11.64, with 69% requiring a 7-day stay or less.

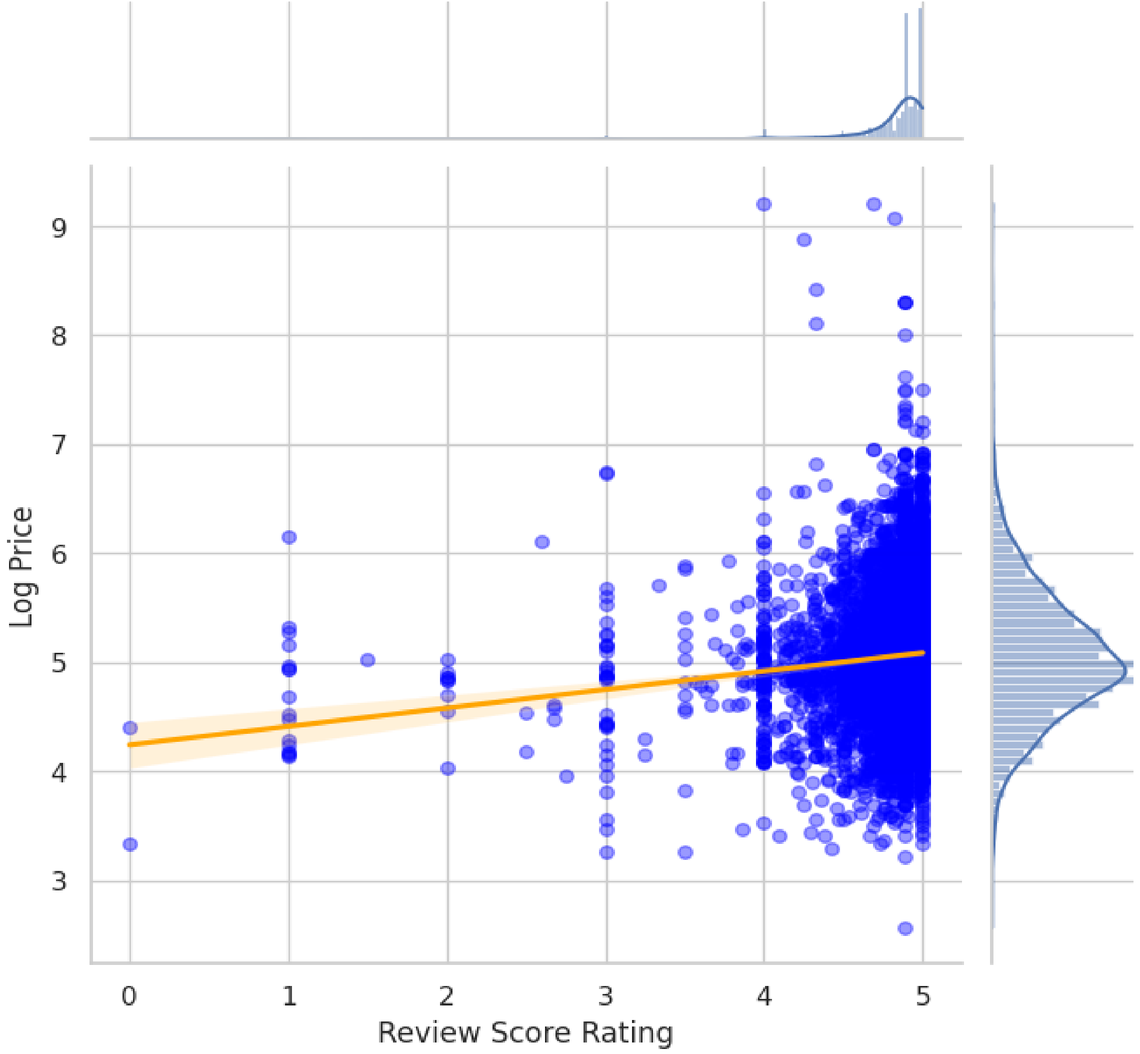
- Barplots show no clear pattern on the left, and a distinct median log price difference for 1 and 7-night stays on the right, indicating the impact of minimum nights on pricing.

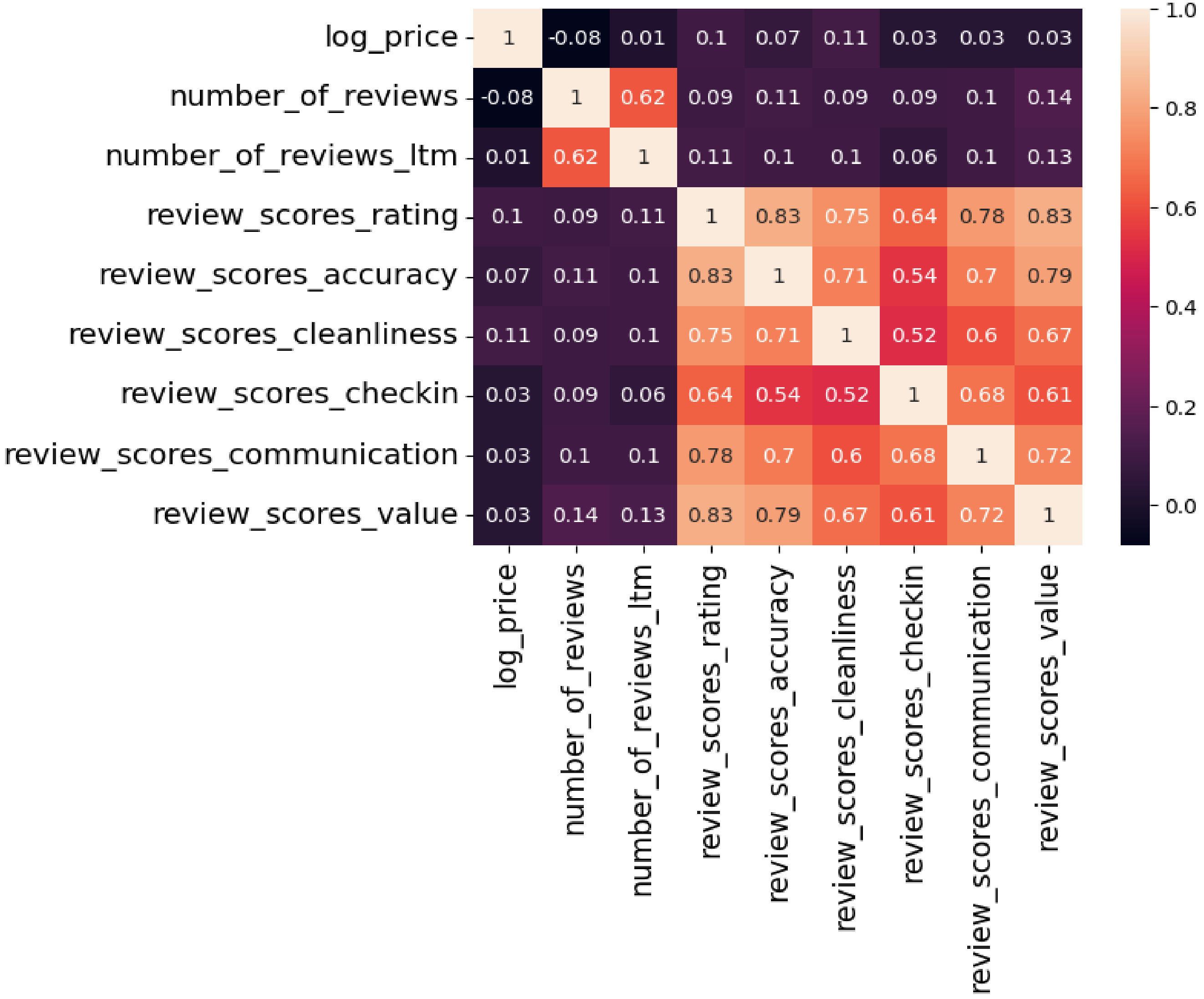






Relationship between Review Score Ratings and Log Price



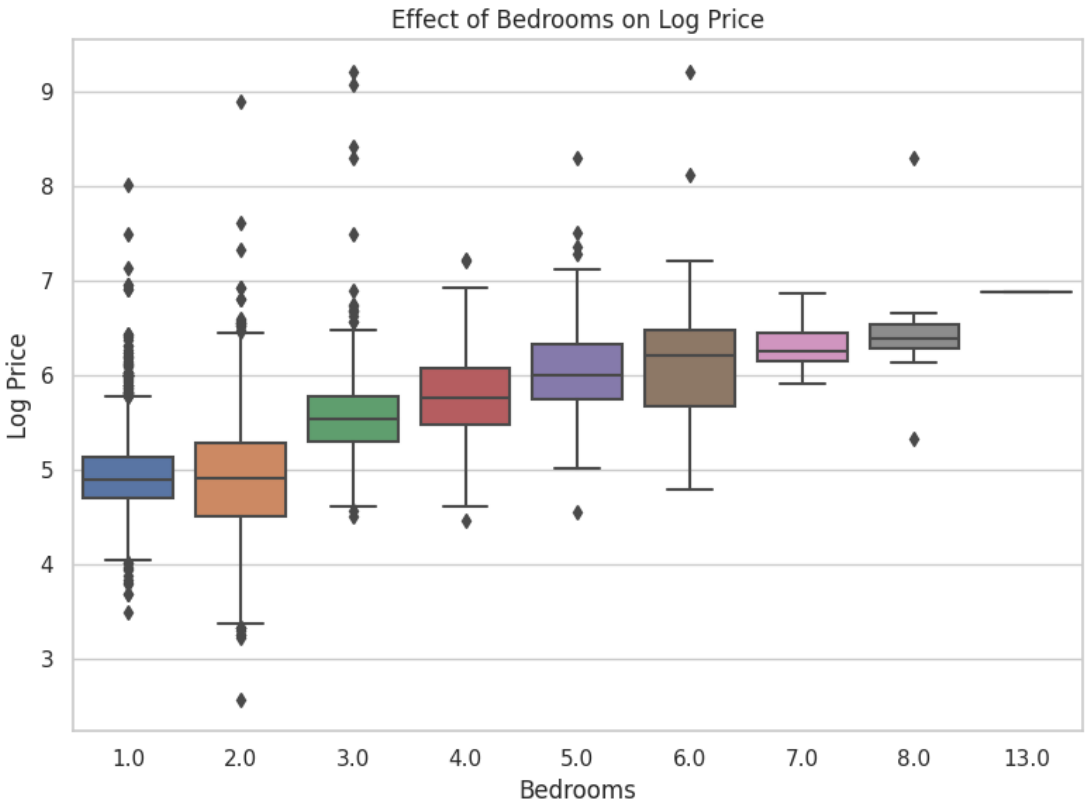


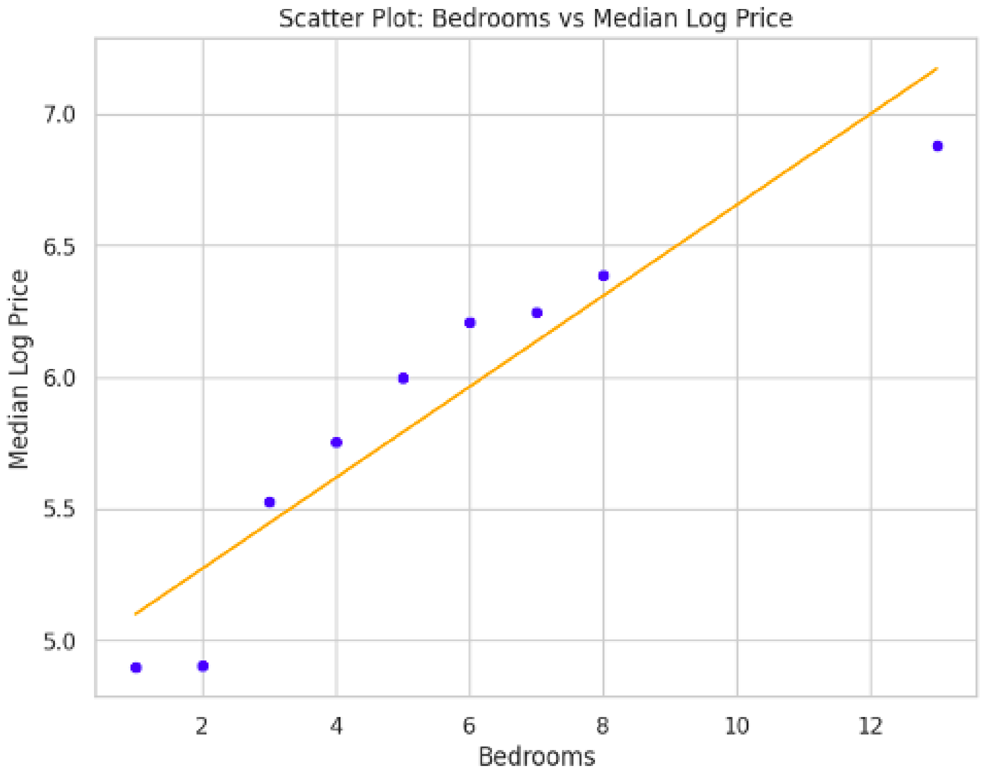
	log_price	number_of_reviews	number_of_reviews_ltm	review_scores_rating	review_scores_accuracy	review_scores_cleanliness	review_scores_checkin	review_scores_communication	review_scores_value
log_price	1	-0.08	0.01	0.1	0.07	0.11	0.03	0.03	0.03
number_of_reviews	-0.08	1	0.62	0.09	0.11	0.09	0.09	0.1	0.14
number_of_reviews_ltm	0.01	0.62	1	0.11	0.1	0.1	0.06	0.1	0.13
review_scores_rating	0.1	0.09	0.11	1	0.83	0.75	0.64	0.78	0.83
review_scores_accuracy	0.07	0.11	0.1	0.83	1	0.71	0.54	0.7	0.79
review_scores_cleanliness	0.11	0.09	0.1	0.75	0.71	1	0.52	0.6	0.67
review_scores_checkin	0.03	0.09	0.06	0.64	0.54	0.52	1	0.68	0.61
review_scores_communication	0.03	0.1	0.1	0.78	0.7	0.6	0.68	1	0.72
review_scores_value	0.03	0.14	0.13	0.83	0.79	0.67	0.61	0.72	1

Bedrooms and Bed

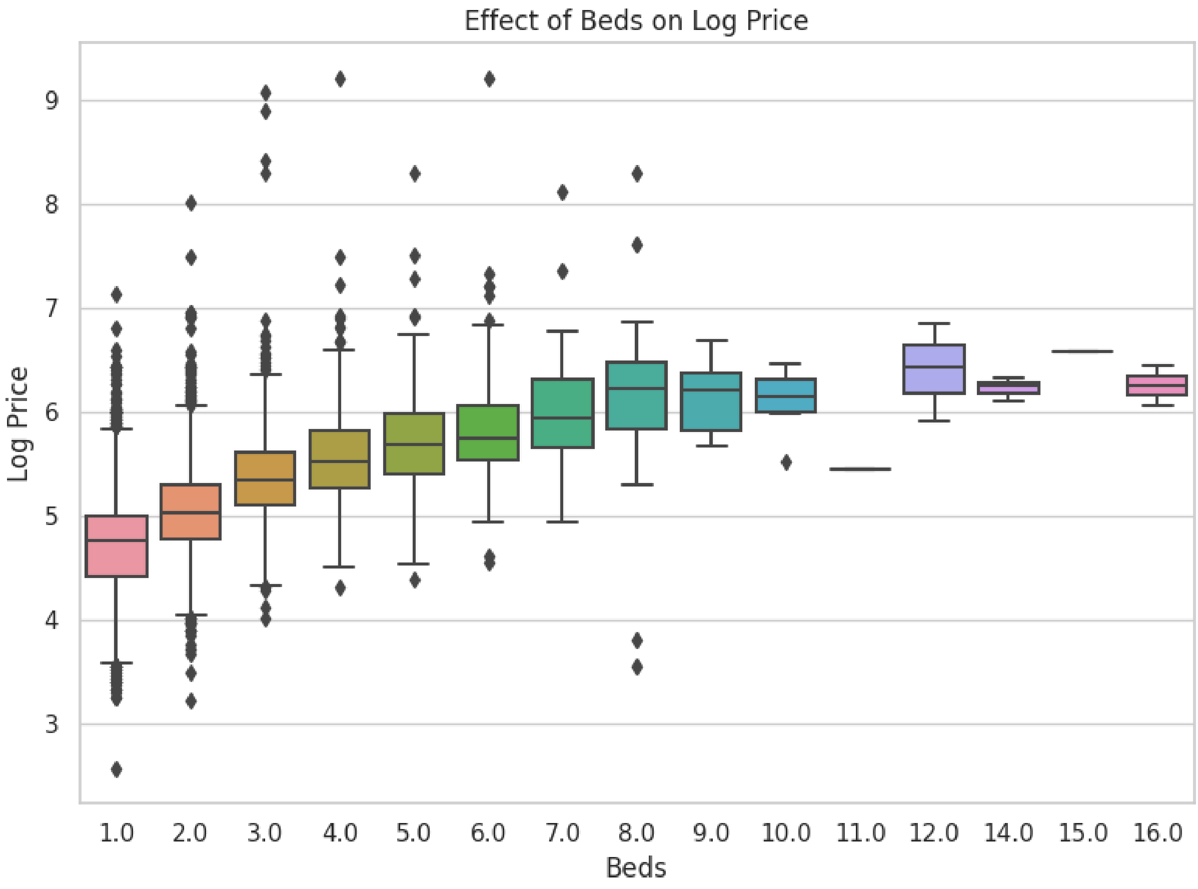
•Scatter plot visually showcases a positive correlation between the number of bedrooms and median log prices.

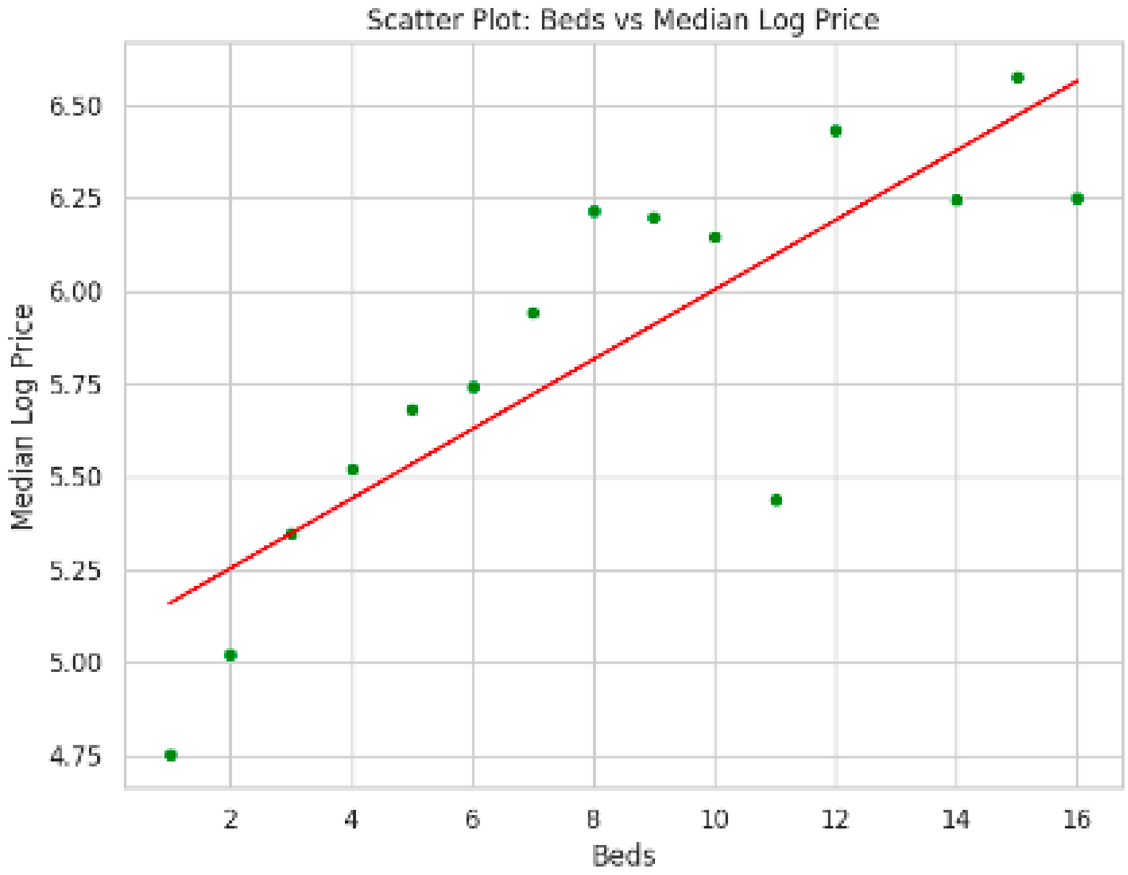
•Orange regression line and accompanying equation quantify the relationship, providing insights into how changes in bedroom count relate to variations in median log prices in the dataset.





Linear Fitting y=0.17x+4.93





Linear Fitting: y=0.09x+5.07

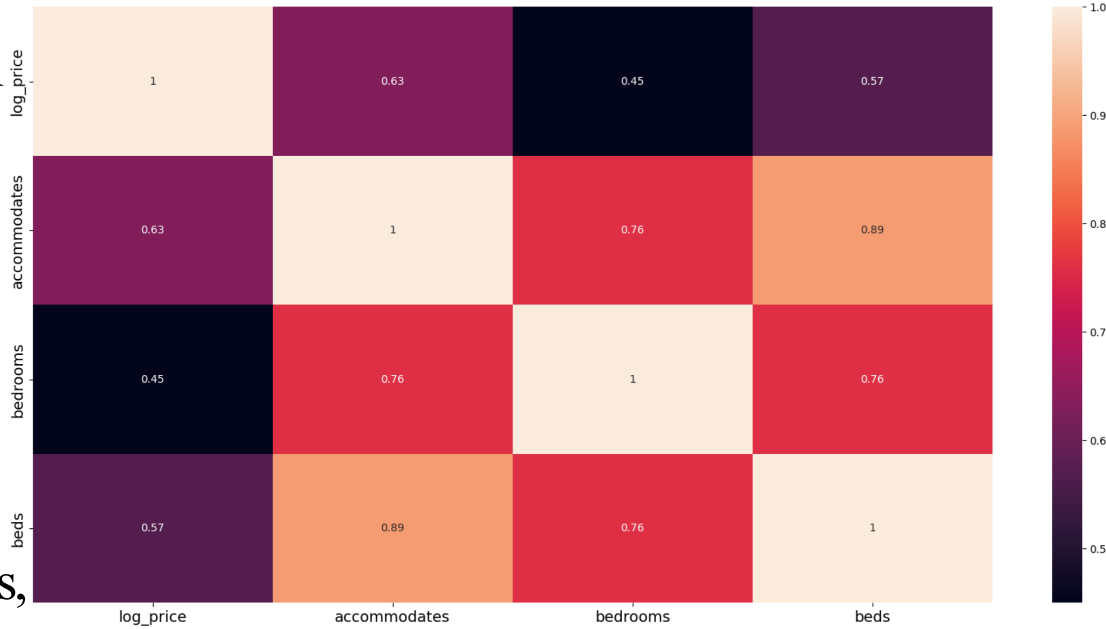
Preprocessing

Categorical Feature Analysis:

- MWU and KW tests assess significance of property_type, room_type, and bed_type in predicting prices.
- Feature engineering and testing reveal significant differences in log price distributions for property_type_simple and room_type.

Dataset Insights:

- Initial examination identifies right-skewed price distribution, normalized through log transformation.
- Correlation analysis connects log prices with numerical features, while exploration of categorical features and amenities uncovers potential predictors and key contributors to pricing dynamics in Airbnb dataset.

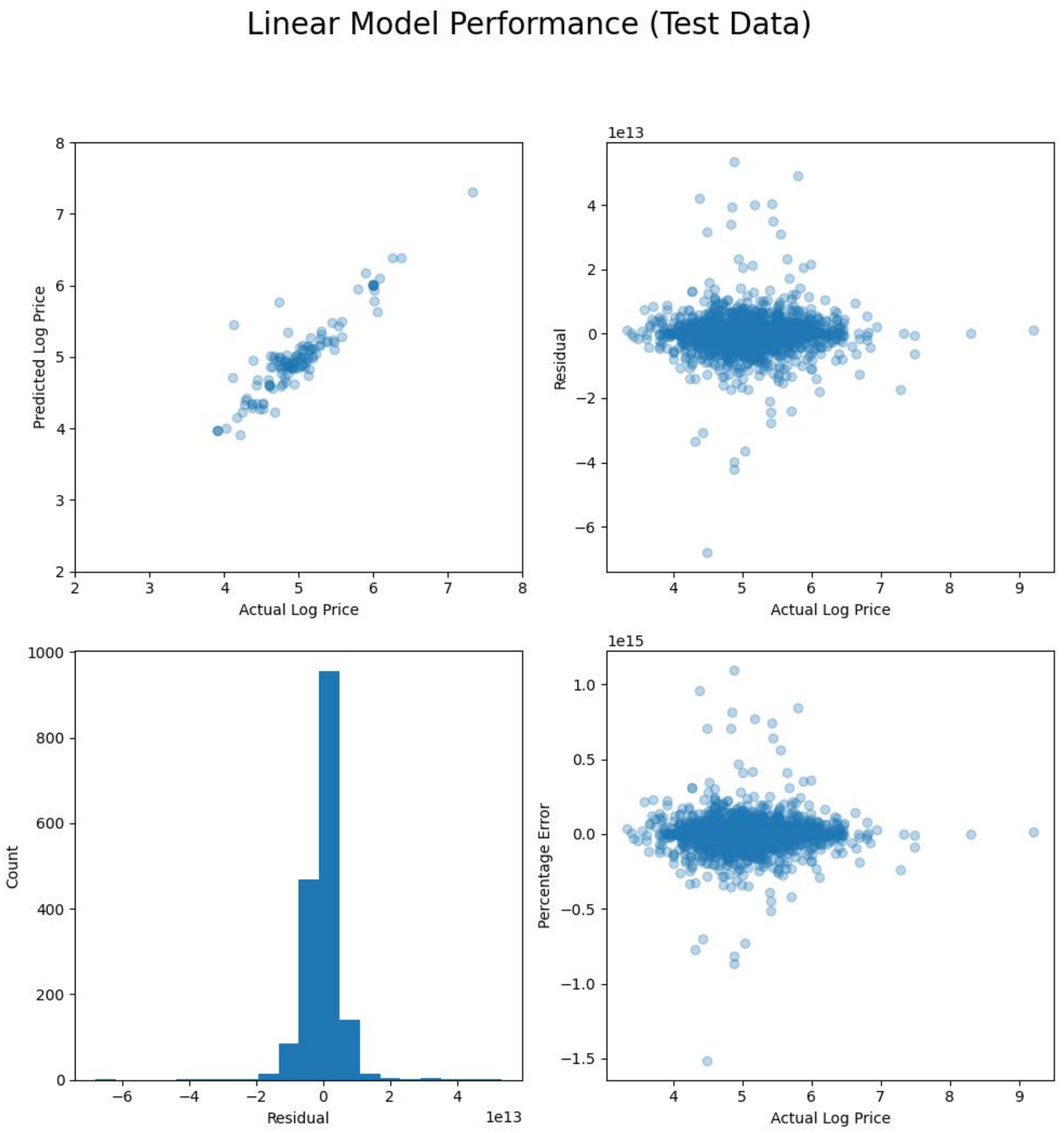


Machine Learning Modeling

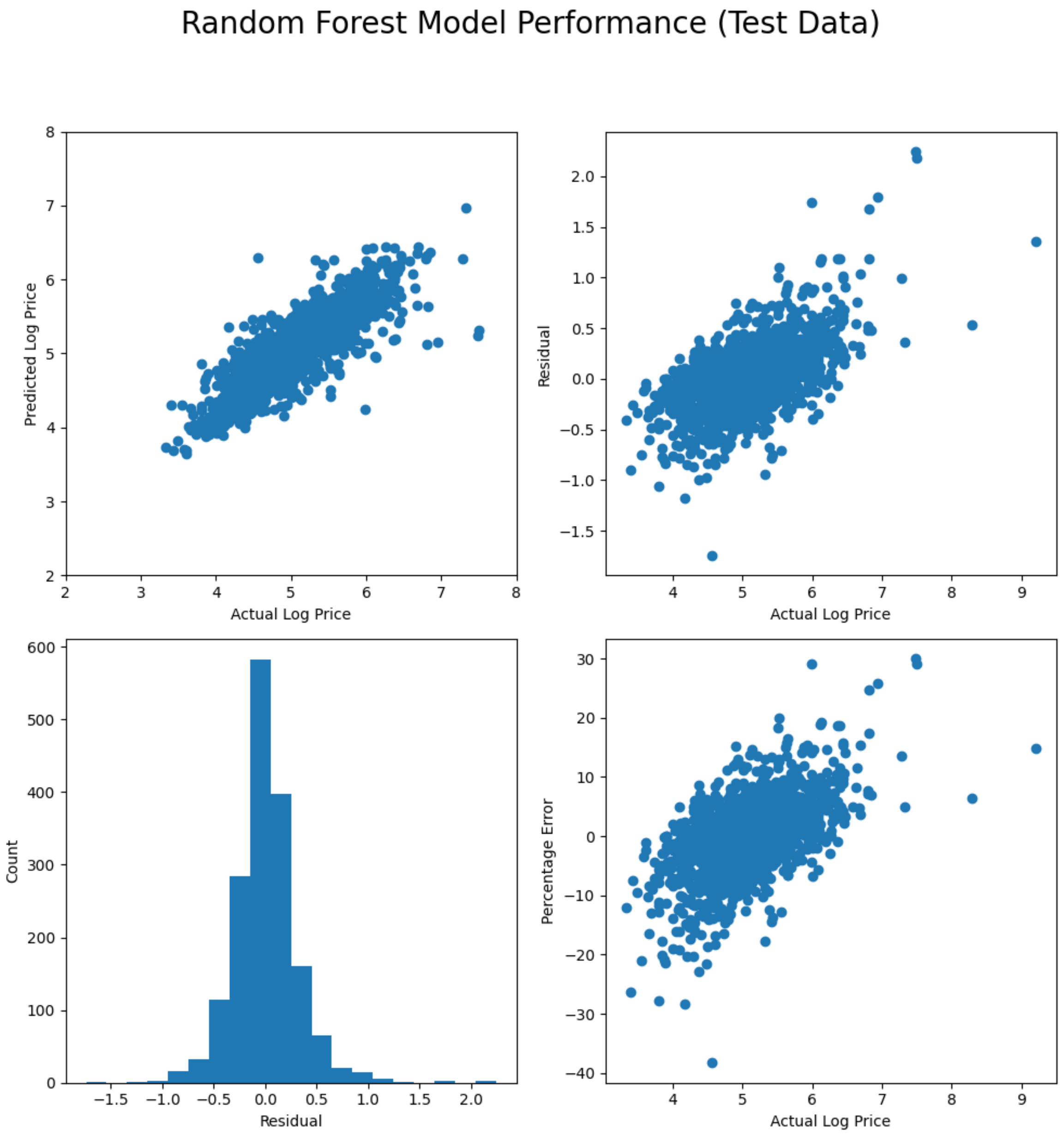
- **Machine Learning Model Development:**
  - Explore linear regression, random forest, gradient boosting, and XGB models with log price as the target variable.
  - Log transformation ensures equal impact on predicting expensive and cheap listings, vital for linear regression.
  - Model selection aligns with dataset characteristics; hyperparameter tuning through grid or randomized search optimizes performance.

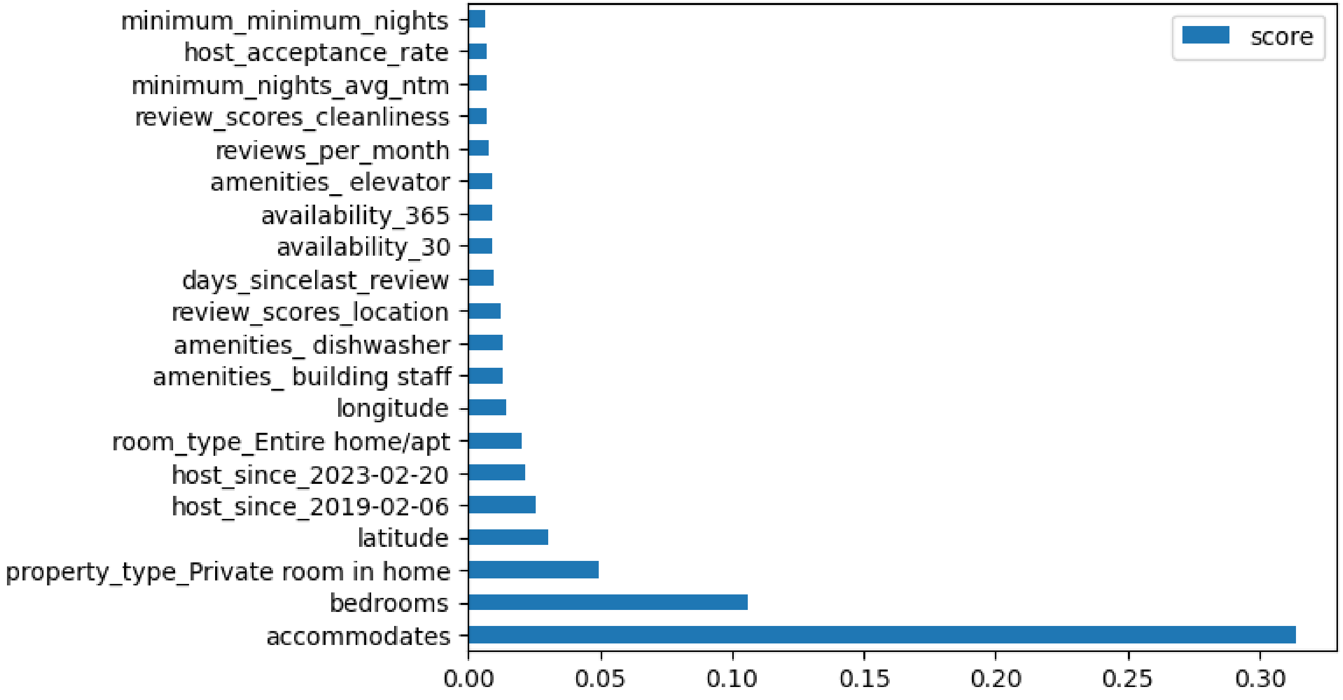
- **Model Training and Evaluation:**
  - Dataset normalization and split into training/testing sets using train_test_split in to the ration 75:25
  - Standardization via StandardScaler prepares data for training and evaluating linear regression.
  - Metrics like mean squared error (MSE) and R-squared (R2) assess model performance.

Linear Regression Model



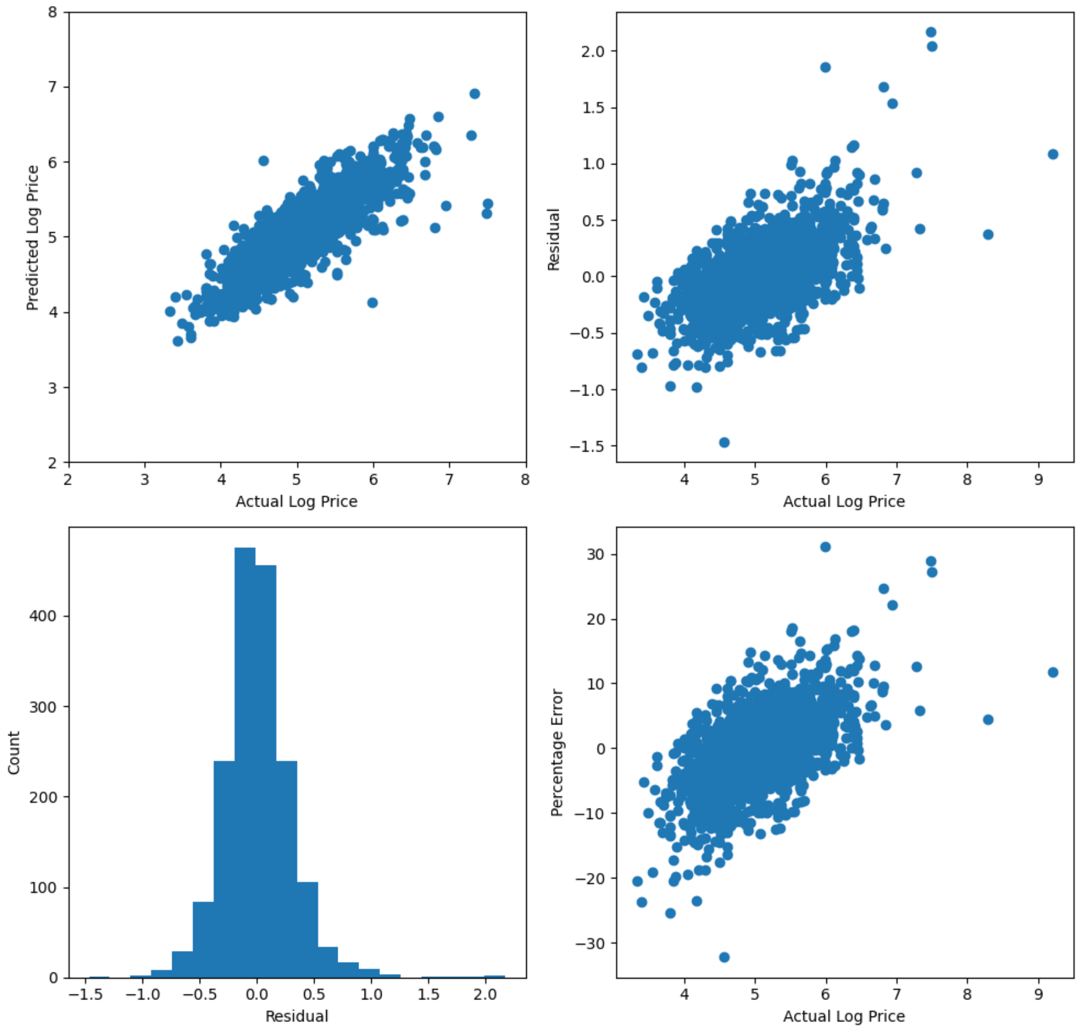
Random Forest

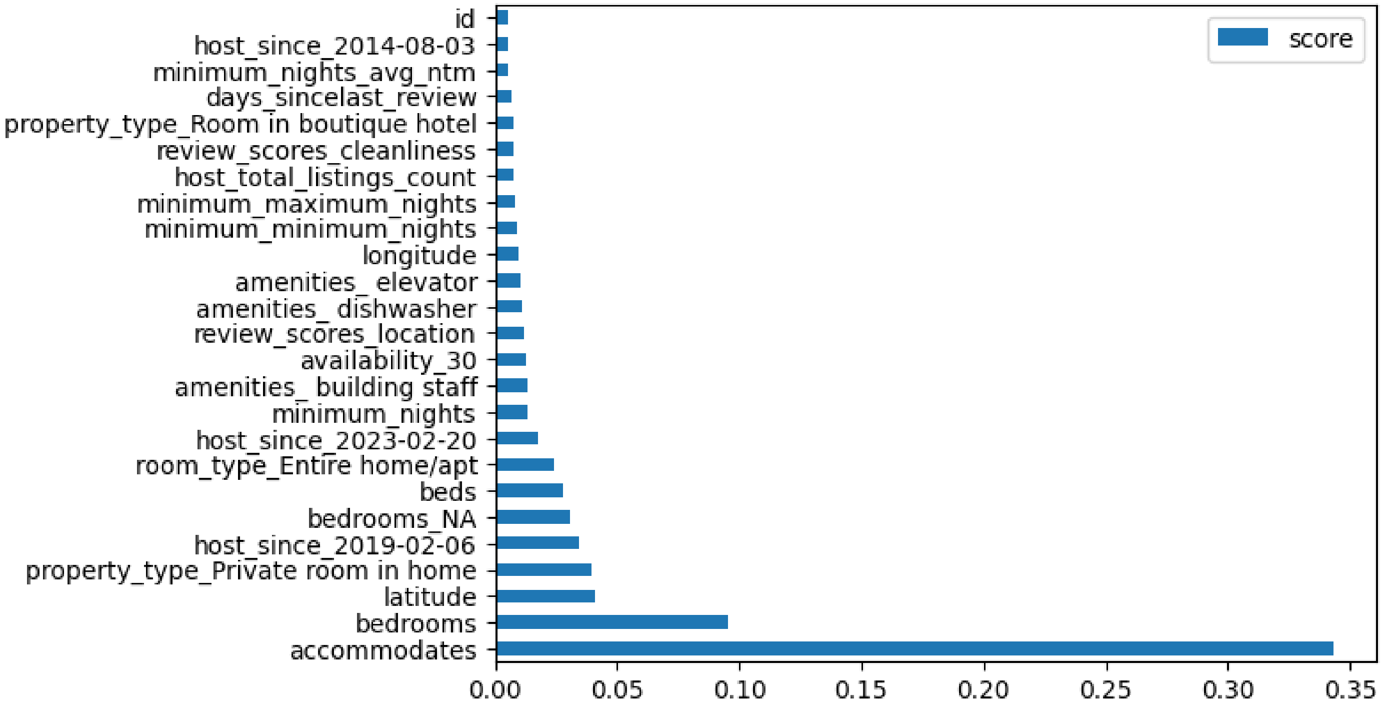




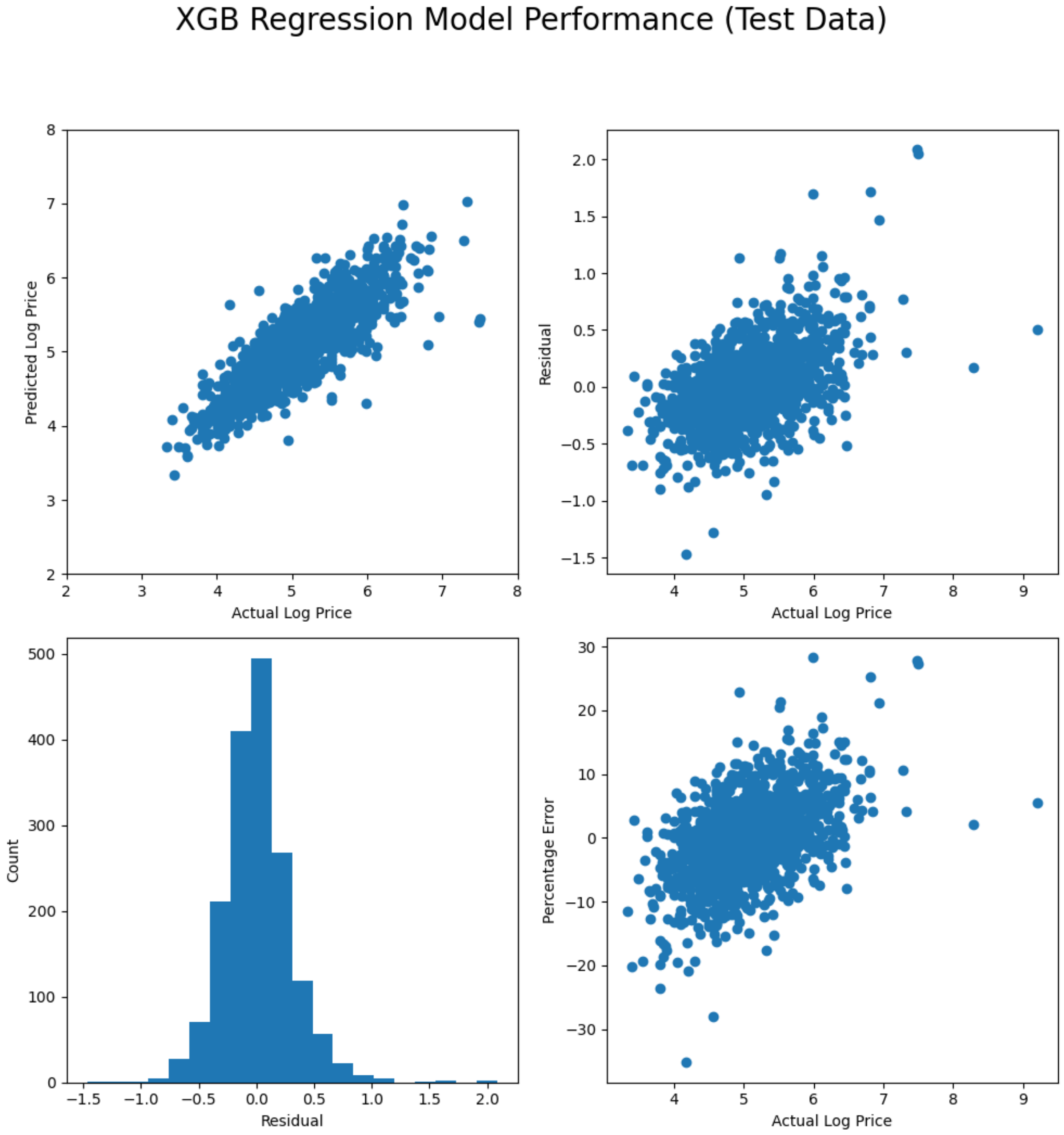
Gradient Boosting

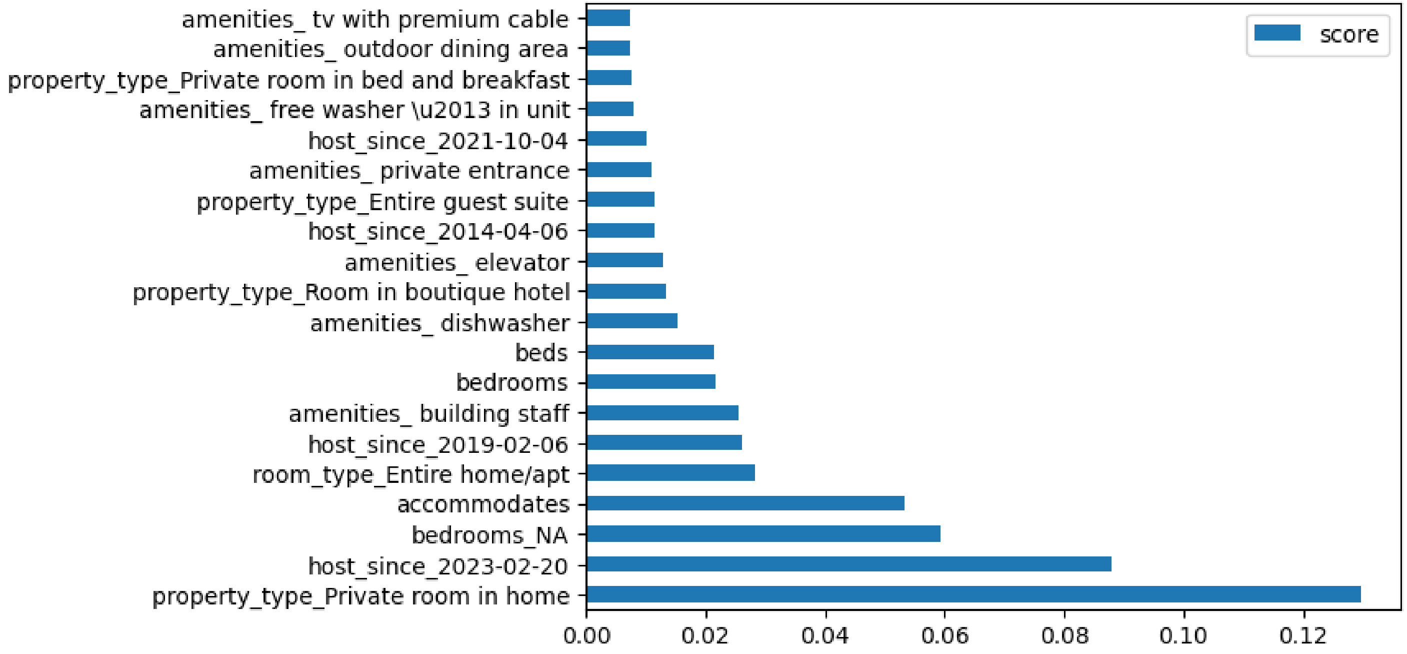
Gradient Boosting Model Performance (Test Data)





XGBoost

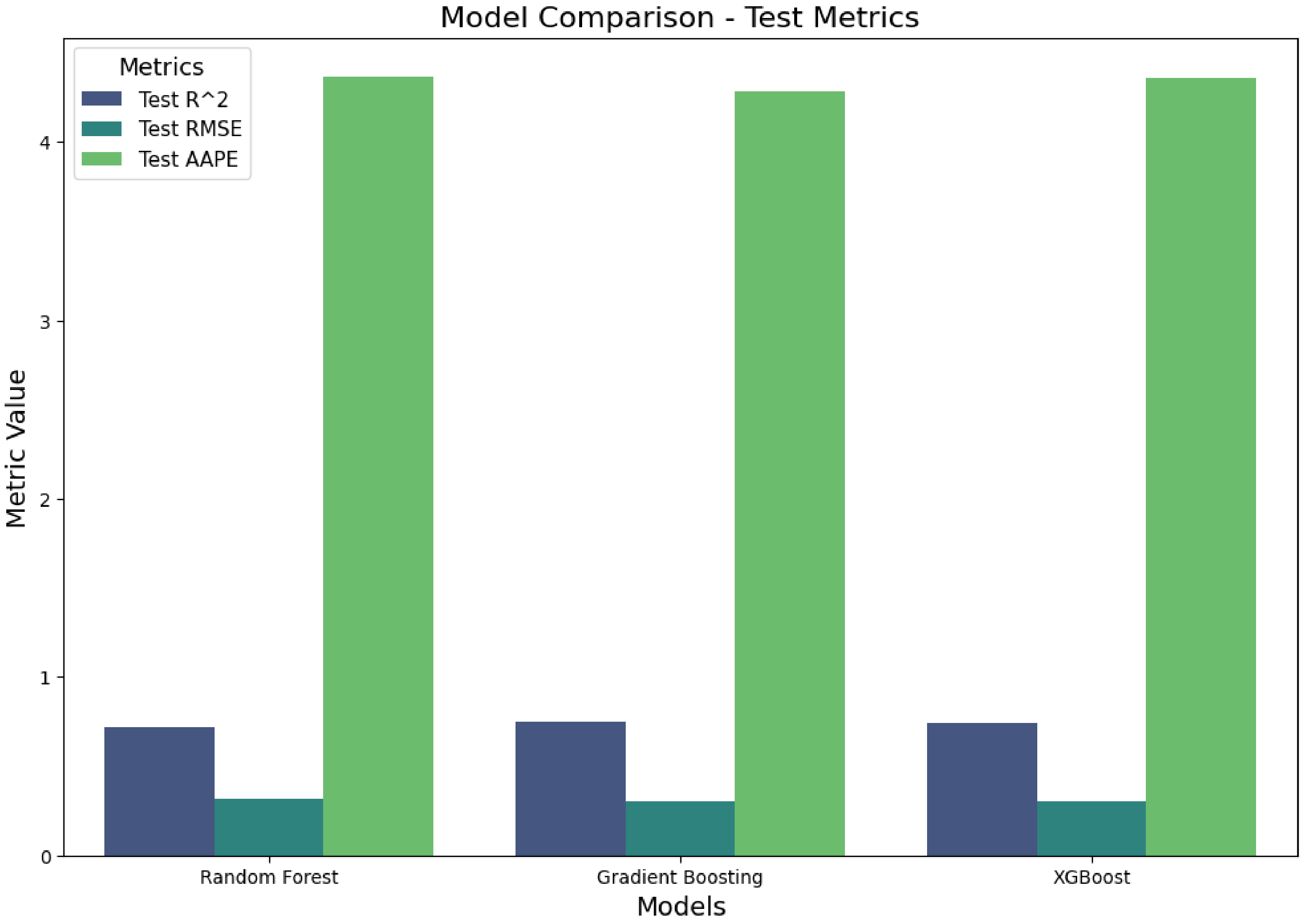




Room type stands out as the most crucial feature, along with the number of accommodates.

Certain amenities like elevator, TV, and gym are also identified as influential factors for accurate price predictions.

Model Comparison



Conclusion

- **Data Science Workflow:**
  - Conducted meticulous EDA, preprocessing, feature engineering, and model evaluation for Seattle Airbnb listings.

- **Model Exploration:**
  - Explored linear regression, random forests, gradient boosting, and XGBoost for capturing intricate patterns.

- **Hyperparameter Tuning:**
  - Optimized models through grid search, focusing on key parameters like learning rate and max tree depth.

- **Model Comparison:**
  - Compared Random Forest, Gradient Boosting, and XGBoost; Gradient Boosting stood with the highest R^2.

- **Best Model:**
  - Identified Gradient Boosting as the superior model, excelling in R^2, RMSE, and AAPE metrics.

- **Results:**
  - The percentage error is less than 15%.
  - The predicted price lies in between $50 to $1100. The approximation error calculation between actual price and predicted price is less than 5%
  - The ideal trend line is expected to fit the predicted price, which strongly supports the model.

Future Recommendations

- **Develop a "Comparative Pricing Tool":** Create a tool that allows hosts to compare their listings with similar properties in their area, including:
    - **Key data points:** Property type, size, amenities, location, and historical booking data.
    - **Price comparison:** Visualize how their current price compares to competitors, offering insights for strategic adjustments.

- **Tailored Modeling and Insights:**
    - **Listing categorization:** Implement a system to categorize listings based on price range ("economy" or "luxury").
    - **Develop separate models:** Train separate machine learning models for each pricing category, potentially leading to more accurate predictions and actionable insights for hosts in different price segments.

- **Enriching the Data:**
    - **Analyze text features:** Extract and analyze textual data from listings and user reviews to understand guest preferences like amenities, location aspects, and unique features.
    - **Refine pricing strategies:** Integrate insights from text analysis into pricing models for more targeted and effective pricing strategies.

Further Research Directions

Dynamic Pricing Analysis: Investigate the dynamic nature of Airbnb pricing, considering factors:

- **Day of the week:** Analyze variations in booking patterns and prices across different days of the week.
- **Holidays and events:** Study the impact of holidays and special events on pricing and demand.
- **Develop adaptive models:** Design models that can adapt to changing market conditions and offer dynamic pricing recommendations.

•**Bridging the Gap Between Prediction and Action:**

- **Price discrepancy analysis:** Identify and analyze discrepancies between actual listed prices and predicted prices from models.
- **Optimize pricing strategies:** Based on the analysis, explore strategies to adjust listed prices for either increased revenue or improved competitiveness.

•**Expanding the Scope:**

- **Explore full property listings:** Investigate the potential impact of offering entire houses or apartments instead of just private rooms or shared spaces.
- **Evaluate home improvements:** Analyze whether specific home improvements (e.g., adding a bedroom) can increase the number of guests accommodated and potentially lead to revenue optimization.

Thank You