Analysis of Airbnb Listings in Marrakesh Price Determinants and Predictive Modeling

Report

1. Introduction

This report presents an analysis of Airbnb listing data in Marrakesh to develop a proxy for real estate supply and demand. The study examines price distributions, identifies key features influencing pricing, handles outliers, analyzes correlations, and implements machine learning models to predict property prices. The insights derived aim to provide valuable market intelligence for stakeholders in Marrakesh's hospitality and real estate sectors.

2. Data Overview and Preparation

The dataset comprises 3,660 Airbnb listings in Marrakesh with 29 initial variables. The data cleaning process revealed several columns with missing values, most notably CoHosts (3,640 missing), LocationDescriptions (1,764 missing), and HouseRules.Aditional (1,504 missing). After cleaning and feature engineering, the dataset expanded to 49 variables through the creation of derived features including price categories, amenity indicators, review metrics, location-based attributes, and title characteristics.

3. Price Distribution Analysis

Price Distribution Characteristics

The price distribution of Airbnb listings in Marrakesh exhibits strong positive skewness. The majority of listings are concentrated in the lower price range (below $500), with a long right tail extending to over $7,000:

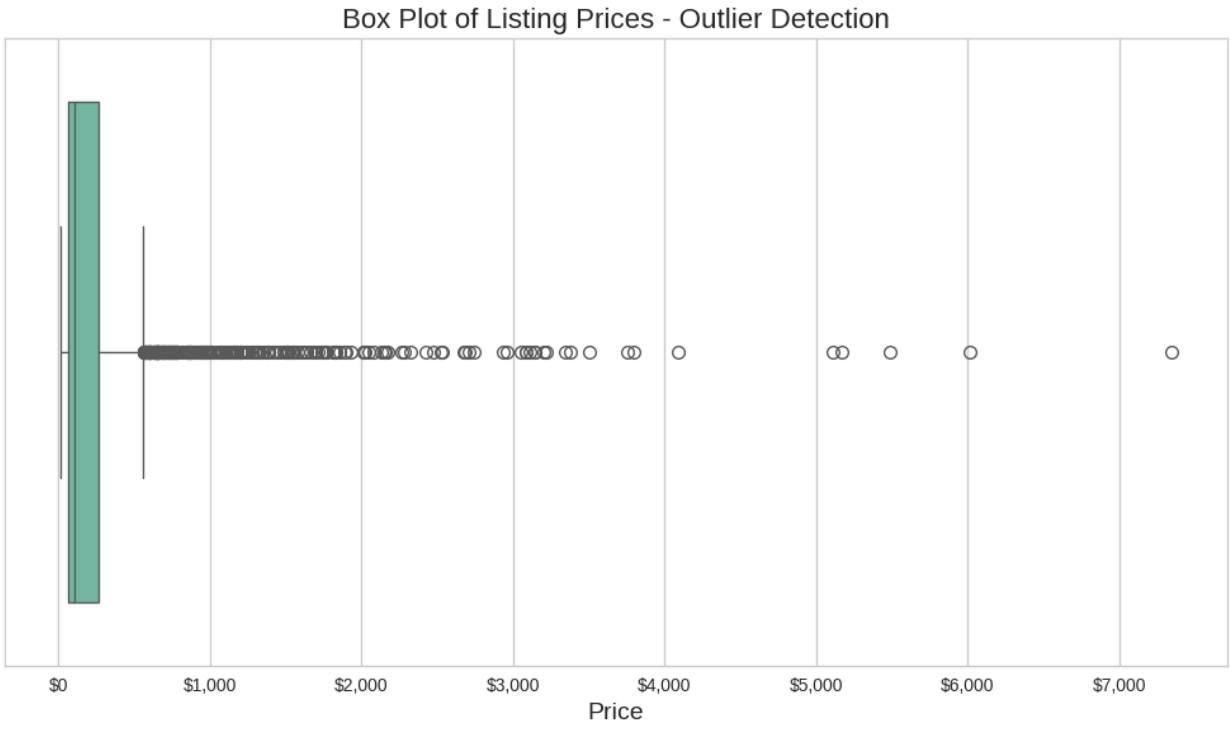
- Mean price: $240.52
- Median price: $107.00
- Minimum price: $12.00
- Maximum price: $7,348.00
- Standard deviation: $416.71

This substantial difference between mean and median prices ($240.52 vs. $107.00) confirms the presence of high-value outliers pulling the mean upward.

Outlier Detection and Processing

The box plot clearly visualizes the presence of price outliers. Using the IQR method, 314 outliers were identified (approximately 8.6% of the dataset), with values exceeding the upper bound of $562.50. These outliers represent premium or luxury properties that command significantly higher prices than typical Marrakesh listings.

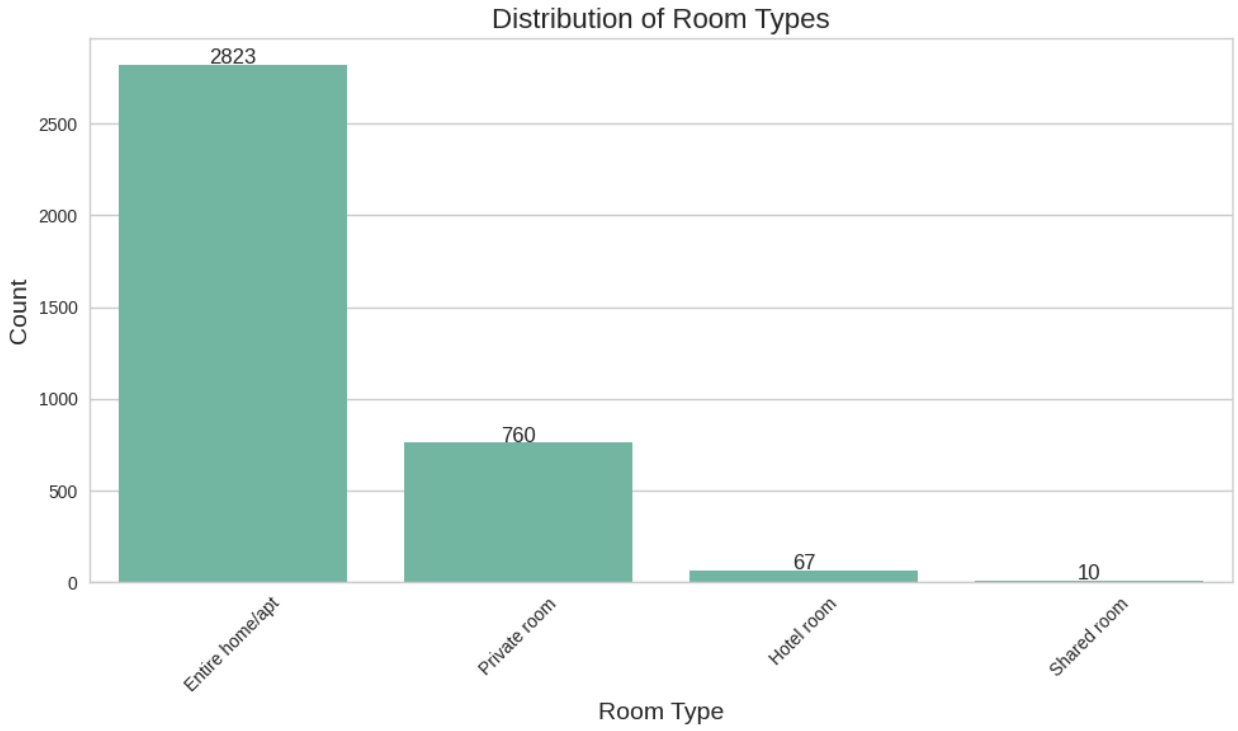
And the effect of outlier processing, comparing the original price distribution (left) with the capped distribution (right). The capping approach preserved the data points while limiting their extreme values, resulting in a more normalized distribution for modeling purposes.



4. Property Characteristics Analysis

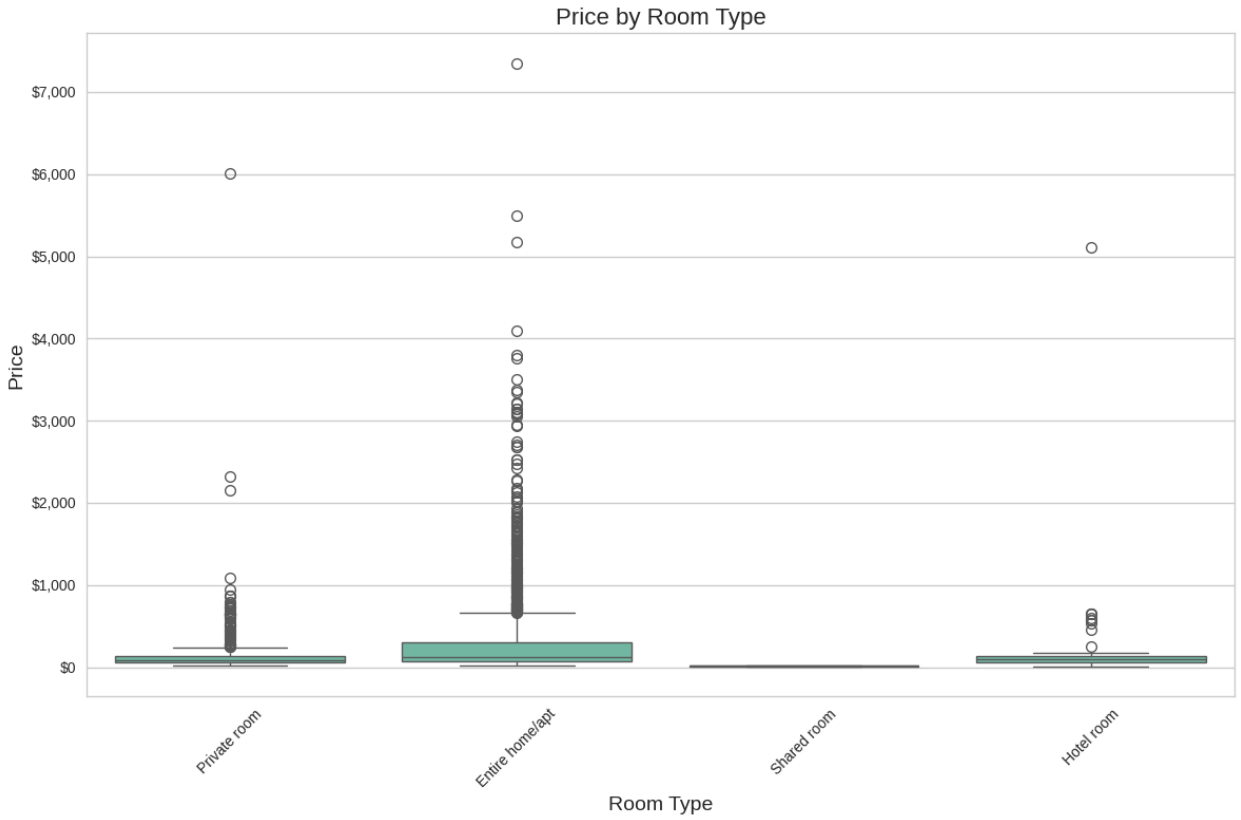
Room Type Distribution

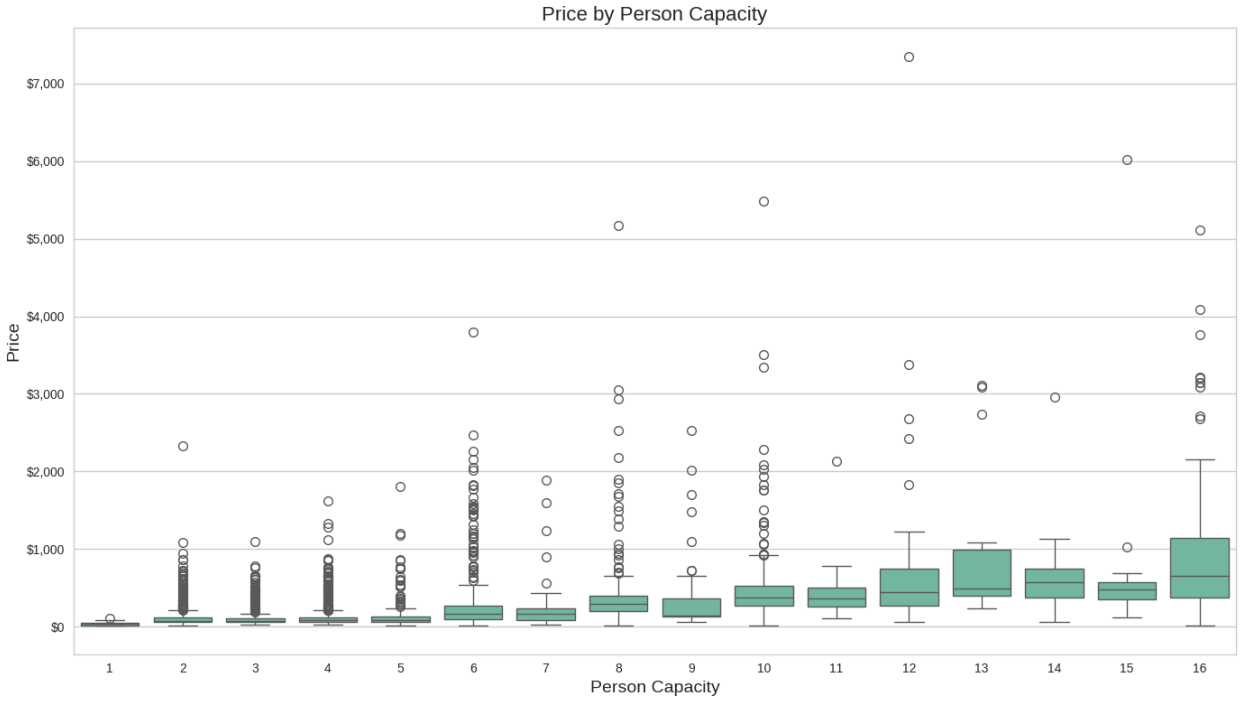
"Entire home/apt" is by far the most common listing type (2,823 listings), followed by "Private room" (760), with "Hotel room" (67) and "Shared room" (10) being relatively rare. This indicates strong market preference for privacy and independence among Marrakesh Airbnb guests.



Price by Room Type

"Entire home/apt" listings demonstrate the widest price range and highest median price, with several high-value outliers reaching up to $7,348. "Private room" and "Hotel room" categories show more modest pricing, while "Shared room" listings have the lowest prices, as expected.

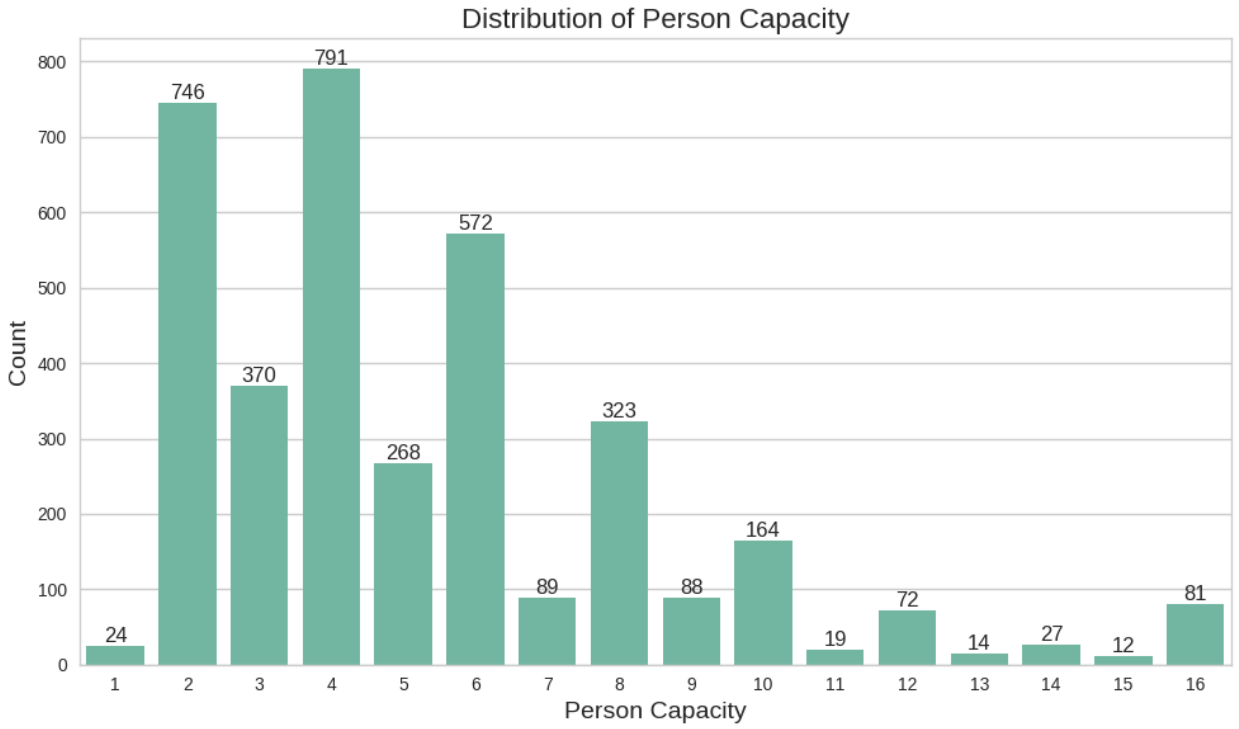




Person Capacity Analysis

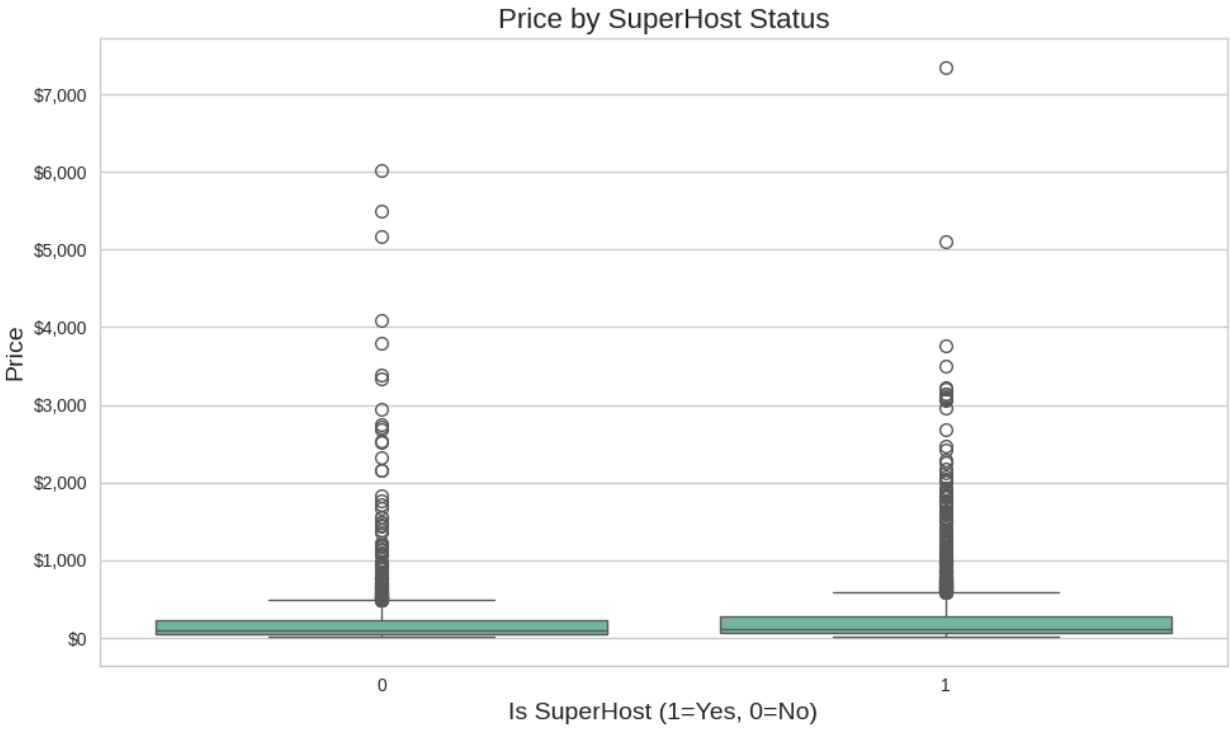
interesting patterns regarding property capacity:

- Most listings accommodate 2-4 guests (with peaks at 2 and 4 persons)
- Larger capacity properties (12-16 persons) are less common but command higher median prices
- A clear positive correlation exists between capacity and price, particularly for properties accommodating 10+ guests
- Properties with capacity for 16 guests show the highest median price and greatest price variation.



SuperHost Status Effect

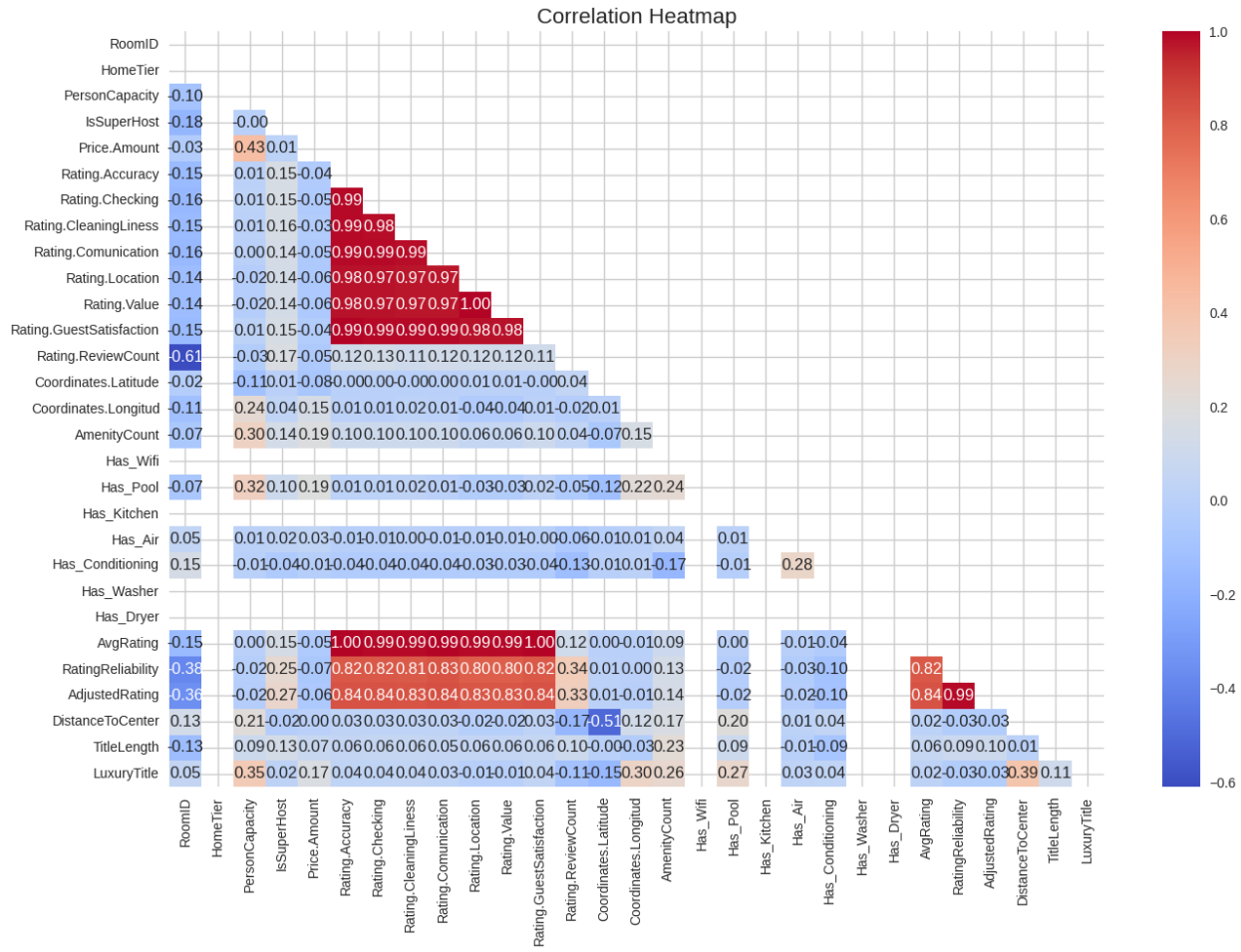
While the price distributions look similar, SuperHost properties (value 1) show slightly higher median prices than non-SuperHost listings (value 0), suggesting a modest price premium associated with SuperHost status.



5. Correlation Analysis

The correlation heatmap reveals several significant price determinants:

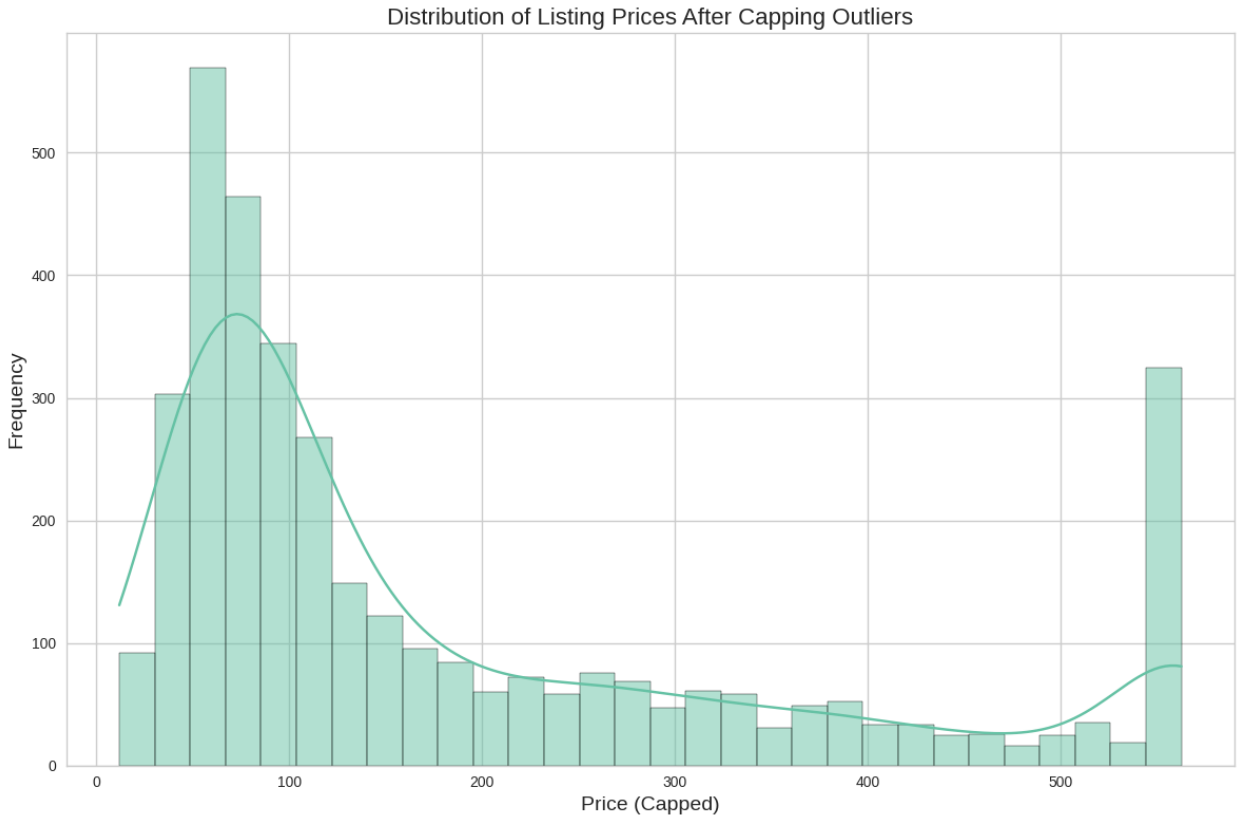
- Strong positive correlation (0.43) between HomeTier and Price.Amount, indicating that higher-tier homes command substantially higher prices
- Moderate positive correlation (0.30) between AmenityCount and price, suggesting that properties with more amenities justify higher pricing
- Notable positive correlation (0.35) between LuxuryTitle and price, showing that listings with luxury-signaling words in titles tend to be priced higher
- Swimming pools (Has_Pool) show positive correlation (0.32) with price, representing a valuable amenity in Marrakesh's climate
- Person capacity shows only weak positive correlation (0.01) with price when considering all listings together (though as seen in Image 5, this relationship becomes stronger for higher-capacity properties)
- Rating metrics show very weak or negligible correlations with price, suggesting that in Marrakesh, physical property attributes impact pricing more than review scores



Original Price Distribution Analysis

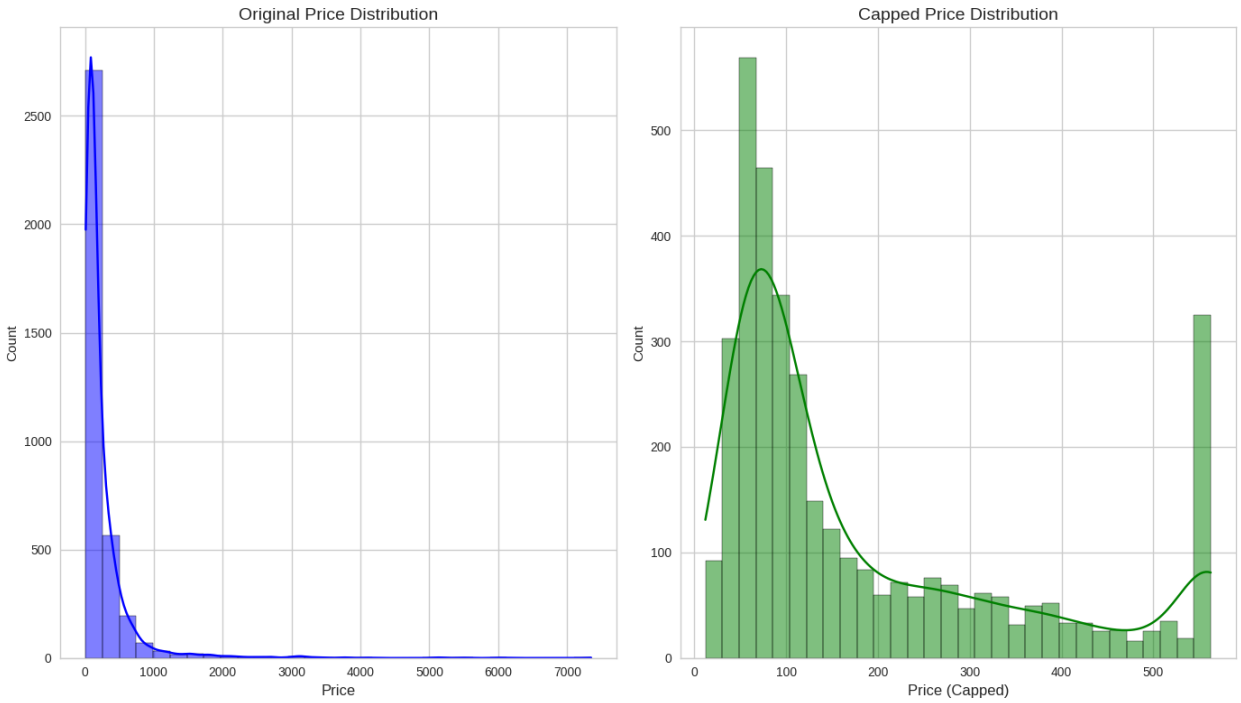
- Extremely right-skewed distribution with a very long tail
- The majority of prices are concentrated at the lower end (below 1000)
- The highest frequency occurs near zero, with approximately 2700-2800 counts
- The scale extends to 7000 on the x-axis
- A blue color scheme is used for visualization
- The distribution follows a power law or exponential decay pattern typical of many economic variables

This distribution suggests a dataset where most items are relatively inexpensive, but there are some very high-priced outliers that stretch the distribution far to the right.



- The x-axis now extends only to about 550 instead of 7000
- A bimodal distribution is evident, with a primary peak around 50-100 and a secondary peak at the maximum cap (around 525)
- The highest frequency is about 550 counts at the first mode
- A green color scheme is used for visualization
- The smooth line suggests this is a kernel density estimate overlaid on the histogram
- There's a long, flat middle section between the two peaks

This capped distribution reveals an artificial "bunching" effect at the maximum price cap, suggesting all values above the cap were set exactly to the cap value. This creates the secondary peak at the right edge of the distribution, which represents all those formerly higher-priced items now grouped at the ceiling price.



6. Predictive Modeling Results

The machine learning experiments tested 17 regression models to predict listing prices. We can observe:

- Tree-based ensemble models significantly outperformed linear models
- ExtraTreesRegressor achieved the best performance with the lowest error metrics:
  - MAE: 0.1186
  - MSE: 0.1752
  - RMSE: 0.3510
  - R^2: 1.0000
- RandomForest and GradientBoosting models also performed well, forming the top three models
- Linear models (Linear Regression, Ridge, Lasso) showed substantially higher error rates

The perfect R^2 score (1.0000) for top models suggests potential overfitting or highly predictable patterns in the dataset after transformation. The extremely low MAPE (Mean Absolute Percentage Error) of 0.0007 for the top model indicates predictions very close to actual values.

7. Key Findings and Business Implications

- Price Segmentation: The Marrakesh Airbnb market shows clear price segmentation, with most properties in the budget to mid-range segment ($65-$264) and a small luxury segment commanding premium prices.

- Property Type Preference: The strong dominance of entire home listings suggests that tourists and visitors to Marrakesh prefer privacy and independence during their stays.
- Capacity Premium: Larger properties accommodating 10+ guests command significantly higher prices, indicating an opportunity in the group travel or multi-family accommodation market.
- Amenity Value: The correlation between amenity count and price highlights the importance of investing in quality amenities, particularly swimming pools, to justify higher rental rates.
- SuperHost Advantage: The modest price premium for SuperHost properties suggests that achieving and maintaining this status provides some pricing power.
- Predictive Power: The high accuracy of tree-based models indicates that physical and categorical property attributes are strong predictors of market price in Marrakesh.

9. Conclusion

This analysis demonstrates that Airbnb listing prices in Marrakesh follow discernible patterns strongly influenced by property characteristics such as home tier, amenities (especially pools), capacity, and property type. The machine learning models, particularly tree-based ensembles, provide accurate price predictions that could serve as a reliable proxy for understanding supply and demand dynamics in Marrakesh's short-term rental market. These insights can guide pricing strategies for hosts, investment decisions for property developers, and market entry strategies for hospitality businesses.