AIRBNB PRICE PREDICTION:

Factors Affecting Short-Term Rental Markets using Machine Learning Models

GROUP 5

- Uyen Nguyen
- Dita Tran
- Katie Nguyen
- My Tang
- Annie Vu

Abstract

This study analyzes the key determinants of Airbnb pricing across 16 major cities in the United States. The analysis is based on a dataset of 213,171 listings with 80 features that describe property characteristics, host attributes, guest reviews, and location information. Four predictive models are applied, including Ordinary Least Squares, Ridge Regression, Lasso Regression, and Random Forest, to evaluate both interpretability and predictive performance. After data cleaning and log transformation of nightly prices, the results indicate that property size, number of bathrooms, number of bedrooms, and amenity availability are the most influential factors in price determination. Location at both the city and neighborhood levels also contributes significantly to price variation. Among all models, Random Forest achieves the highest predictive performance with a test R^2 of 0.756. While the primary focus is on listing-level determinants, the study also provides extended insights into market structure and how the differences of supply and demand across cities influence pricing patterns. In addition, the potential impacts of major tourism events, such as the 2026 FIFA World Cup, are discussed to demonstrate how large demand surges may temporarily influence pricing. Future research can expand on these results by incorporating event-based data and policy indicators to better capture dynamic changes in short-term rental markets.

Keywords: Airbnb pricing, machine learning, hedonic analysis, market structure, supply and demand dynamics, tourism impacts

I. Introduction

The rapid growth of the sharing economy has reshaped the hospitality industry by allowing individuals to provide short-term accommodation through digital platforms. Airbnb is the largest and most influential of these platforms, offering a wide range of rental options that vary in size, quality, and location. Unlike hotels, which typically use standardized pricing strategies, Airbnb hosts can adjust prices freely based on their own understanding of property value, local competition, and guest demand. As a result, pricing outcomes are highly diverse even within the same city, making it important to identify the factors that determine what guests are willing to pay.

Previous studies show that Airbnb pricing depends on several groups of attributes, including property characteristics, host credibility, review quality, and neighborhood desirability. However, much of the existing literature examines these elements separately, and less attention has been paid to the combined and interacting effects of these features in large and diverse urban markets. In addition, pricing does not occur in isolation. City-level supply and demand conditions play a major role in shaping price competition and price inequality. Some cities have an oversupply of Airbnb listings, whereas others have more limited availability and tend to remain at more stable pricing. Understanding these broader

market patterns can provide more practical insights for hosts, guests, and policymakers.

External tourism demand shocks also have the potential to disrupt normal pricing behavior. For example, the 2026 FIFA World Cup is expected to generate a sharp increase in travel demand across several host cities in the United States. This type of large events may create temporary surges in prices, along with rapid changes in supply as more residents enter the short-term rental market. Few studies have evaluated how such events influence Airbnb pricing or whether predictive models remain reliable under unusual conditions.

Based on these research gaps, this study examines Airbnb pricing using housing, hospitality, and data-driven perspectives. The goal is to identify the main determinants of price, evaluate model performance across various predictive techniques, and provide further insights into how market structure and external tourism pressure influence price variation. To address these objectives, we apply four regression and machine learning models to a large dataset of Airbnb listings from 16 major U.S. cities.

II. Literature Review

The rapid expansion of the sharing economy has transformed traditional hospitality markets, with Airbnb emerging as the dominant peer-to-peer accommodation platform. Research on Airbnb has expanded rapidly, drawing from hospitality economics, urban studies, consumer behavior, and data science. Unlike hotels, which rely on standardized pricing models, Airbnb hosts set their own nightly rates based on their understanding of property features, local market conditions, and guest expectations. As a result, understanding the determinants of Airbnb prices has become a central topic in contemporary tourism and housing research. We synthesized existing findings into five major themes: pricing determinants, locational attributes, host reputation and trust, review and sentiment effects, and machine learning–based price prediction.

Regarding pricing determinants, foundational work applies hedonic or implicit market models to quantify how listing characteristics influence nightly rates. Traditional hospitality studies provide an important baseline. Chen and Rothschild (2010) and Carvell and Herrin (1990) show that hotel room prices reflect structural attributes such as room size, amenities, accessibility, and perceived quality. Subsequent Airbnb-specific studies adopt similar frameworks but incorporate platform-specific attributes. Teubner et al. (2017) identify five broad categories shaping Airbnb prices—property characteristics, host reputation, rental rules, location, and listing visibility—and find that property-level features such as room type, amenities, and maximum occupancy significantly increase prices. Zhang, Ye, and Law (2011) further support the argument that accommodation characteristics follow a “hierarchy of needs,” where amenities and comfort-level factors command higher willingness to pay. Jiao and Bai’s (2023) empirical analysis across forty U.S. cities reinforces the strong predictive value of structural

listing variables such as sleeps, property type, and number of bathrooms. Complementing these hedonic results, Guttentag and Smith (2017) argue that Airbnb competes directly with hotels by offering similar performance expectations on core attributes such as cleanliness and comfort, which strengthens the role of property-level determinants in pricing. Collectively, this stream of research confirms that structural listing features such as accommodates, bedrooms, bathrooms, beds, amenities, and room type form the foundational layer of price-setting on Airbnb.

In terms of locational attributes, we find a consistent finding across the literature is that location exerts one of the strongest influences on Airbnb pricing. Deboosere et al. (2019) conduct a multilevel hedonic analysis and show that neighborhood characteristics such as density, transit accessibility, land use, and local amenities, substantially shape both prices and host revenue. Their study highlights that even after controlling for property-level features, spatial context remains a major driver of price heterogeneity. Urban housing research supports these dynamics at the broader market level. Barron, Kung, and Proserpio (2017) demonstrate that the spatial concentration of Airbnb listings affects local housing affordability, suggesting that neighborhoods with higher tourism demand and supply pressures see elevated rental values. Similarly, Horn and Merante (2017) find that Airbnb activity in Boston is correlated with higher long-term rents, emphasizing the extent to which neighborhood desirability intersects with short-term rental pricing. Policy-focused analyses, such as Wachsmuth et al. (2017), show that Airbnb reshapes urban spatial patterns by encouraging listings in central or high-demand neighborhoods, where hosts can extract higher revenues. These findings collectively underline the importance of incorporating city, neighbourhood labels, and location density into pricing models, as neighborhood context interacts with property quality and demand intensity to produce wide variation in nightly prices.

In addition, because Airbnb transactions occur in a digitally mediated peer-to-peer marketplace, trust and reputation mechanisms are crucial for reducing perceived risk. The literature consistently shows that host credibility significantly affects pricing. Ert, Fleischer, and Magen (2016) provide one of the earliest behavioral insights: hosts' profile photos serve as powerful trust signals, and more trustworthy images lead to significantly higher booking likelihood and willingness to pay. Their findings demonstrate that reputation extends beyond numerical ratings to include visual and relational cues. Lee et al. (2015) similarly show that social features are strongly associated with room sales and booking performance. Teubner et al. (2017) reinforce these patterns by demonstrating that host reputation metrics (Superhost status, reviews, identity verification) carry price premiums even after controlling for property characteristics. Across these studies, host attributes function as quality assurance indicators that reduce uncertainty for guests, and as a result, hosts with stronger trust signals can charge higher nightly rates. These findings justify the

inclusion of response rate, acceptance rate, Superhost status, and identity verification as key predictors of price.

On the other hand, guest reviews represent one of the most important trust and value signals on Airbnb. A growing body of literature demonstrates that both numeric rating components and textual review sentiment influence price. Almeida, Nunes, and Machado (2025) use a hedonic pricing model to show that higher review scores, across cleanliness, check-in, communication, and overall rating, predict higher nightly prices. The authors argue that ratings act as a cumulative measure of guest satisfaction and perceived listing quality. Lin and Yang (2024) similarly emphasize the role of review scores in shaping short-term housing prices, showing that highly rated listings consistently receive higher demand and higher pricing across 26 U.S. regions. Recent work integrates sentiment analysis to capture more nuanced dimensions of customer feedback. Kalehbasti, Singh, and Rao (2021) demonstrate that sentiment extracted from review text significantly improves price prediction accuracy in machine learning models, suggesting that the emotional tone of reviews conveys information beyond numerical scores. Chapman et al. (2023) arrive at similar conclusions, showing that sentiment-enhanced models outperform those relying solely on structured rating variables. This body of evidence strongly supports including review_scores_rating, sub-score variables, and sentiment metrics as components of price modeling.

Lastly, while we notice that hedonic pricing models provide interpretability, recent research increasingly relies on machine learning (ML) approaches to capture nonlinear relationships and interactions among Airbnb features. ML models are especially useful when working with high-dimensional, heterogeneous data involving property characteristics, location, host attributes, and review text. Medpalliwar et al. (2023) build a machine learning framework for Airbnb price prediction using algorithms such as Random Forest and Gradient Boosting, finding that tree-based ensemble models outperform linear regression. Alharbi (2023) strengthens this insight by combining structured listing features with sentiment analysis and concluding that ML models yield higher accuracy and better generalizability than traditional hedonic approaches. Camatti et al. (2024) provide a comprehensive comparison between “classic” econometric models and artificial intelligence techniques, showing that ML approaches, especially Random Forests and neural networks, consistently achieve lower prediction errors. Chapman et al. (2023) likewise emphasize the superiority of ML models for dynamic short-term rental markets. Across these studies, Random Forest repeatedly emerges as a strong performer, capable of modeling complex interactions without overfitting. This supports the methodological choice in the present study to compare linear models (OLS, Lasso, Ridge) with ensemble models such as Random Forest for price prediction.

The literature consistently highlights the multidimensional nature of Airbnb pricing. Structural listing characteristics form the baseline determinants of price, but neighborhood context, host reputation, review quality, and guest sentiment all exert substantial influence. Machine learning models have proven particularly effective at capturing the interactions among these variables. Together, these insights provide the conceptual and empirical foundation for our study's modeling approach and feature selection.

III. Data

A. Data Overview

The data for this study comes from InsideAirbnb, a public dataset of Airbnb listings, using the 2025 release. The dataset contains 213,171 observations and 80 variables. These variables describe listing characteristics, host attributes, guest demand, and various metadata fields.

The variables in the dataset can be grouped into several conceptual categories. The target variable of interest is the nightly price. Listing capacity features include accommodates, bedrooms, bathrooms, and beds. Demand and rating variables include number_of_reviews, reviews_per_month, and review_scores_rating. Host quality indicators include host_is_superhost, host_response_rate, and host_acceptance_rate. The dataset also contains categorical features such as city, neighbourhood_cleansed, room_type, and property_type. Several additional fields capture amenities, availability, and descriptive metadata such as host_about and neighbourhood_overview

A missing-value analysis of the raw dataset shows substantial variation in data completeness. Several numerical fields have more than 40,000 missing entries, such as reviews_per_month (47,166 missing), price (44,597 missing), and beds (44,133 missing).. These missing values highlight the need for careful preprocessing prior to modeling.

```
Missing Values:
calendar_updated              213171
license                       117875
neighborhood_overview         102182
neighbourhood                 102179
neighbourhood_group_cleansed   88033
host_about                     84147
host_location                  48369
review_scores_value            47235
review_scores_location         47235
review_scores_checkin          47231
review_scores_accuracy         47216
review_scores_communication    47211
review_scores_cleanliness      47204
first_review                   47166
last_review                    47166
review_scores_rating           47166
reviews_per_month              47166
price                          44597
estimated_revenue_l365d        44597
beds                           44133
```

Figure 1. Missing Values Count

Summary statistics for the main numerical variables are shown in Table 2. Review-related variables also exhibit substantial variability, reflecting differing levels of guest activity across listings. Capacity variables such as accommodates and bedrooms show wide ranges, indicating a mix of small private rooms and large multi-bedroom properties.

B. Data Cleaning and Pre-Processing

1. Price cleaning

The original price variable was stored as a string (e.g., "$120.00"), so we first standardized it by stripping currency symbols and commas and converting values to floating-point numbers (e.g.,

"$120.00" → 120.0). The distribution of nightly prices is highly right-skewed, with a mean of approximately $502 and a median of $287. Therefore, we applied a natural logarithm transformation to price to stabilize variance, reduce skewness, and improve interpretability. All subsequent models use log(price) as the dependent variable.

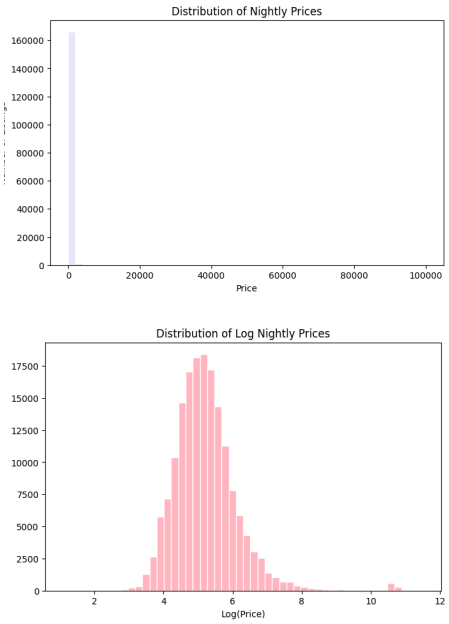

Figure 2. Distribution of Price before vs. after Log Transformation

2. Missing Values and Variables Transformation

To prepare the dataset for analysis, we cleaned and standardized several categories of variables. Host indicators such as host_is_superhost were converted into binary form, with missing values treated as non-superhost, while host response and acceptance rates were cleaned and transformed into numerical percentages. Review scores, review counts, and listing capacity features (e.g., accommodates, bedrooms, bathrooms, beds) were also converted to numeric format and imputed using median values to handle occasional missing entries.

The amenities field was simplified by extracting an overall amenities_count and creating indicators for the most common amenities across all cities. To reduce sparsity in location categories, only the top five neighbourhoods per city were retained as individual categories, and all others were grouped into an "Other" label. Finally, key categorical variables - including city, room type, property type, and neighbourhood - were one-hot encoded to create the final modeling dataset.

C. Explanatory Data Analysis

1. Market Structure

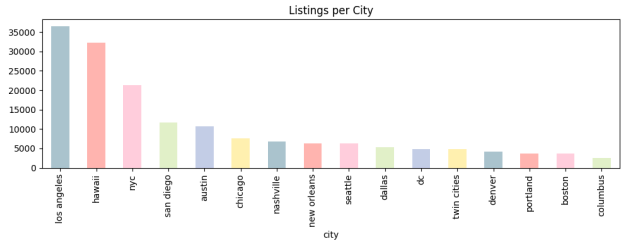

Figure 3. Listings per City

The distribution of listings across cities shows large differences in market size. Cities like Los Angeles,

Hawaii, and New York have the highest number of listings, which suggests they are major travel and tourism hubs. In contrast, cities such as Columbus, Boston, and Portland have much fewer listings. This uneven supply is important because markets with more listings tend to have stronger competition and more price variation, while smaller markets may have more stable or lower average prices.

2. Correlation between numeric predictors

Figure 4. Correlation Heatmap between Numeric Predictors

The correlation heatmap reveals several clear patterns among the numerical predictors. Capacity-related variables, such as accommodates, bedrooms, bathrooms, and beds, are strongly correlated, showing how larger listings tend to have more rooms and higher prices. Review score variables also cluster together, suggesting they capture similar aspects of listing quality. Host quality variables show weaker relationships with price, implying that response or acceptance rates play a smaller role in pricing decisions.

3. Listing Types vs Price

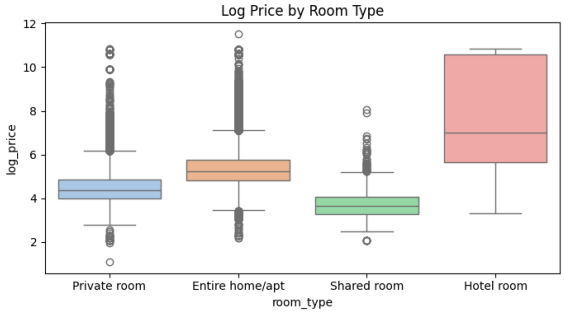

Figure 5. Log Price by Room Type

Room type strongly influences pricing. The Log Price by Room Type chart shows a clear hierarchy: Entire homes/apartments and hotel-style rooms have the highest median log-prices, private rooms fall in the middle, and shared rooms remain the most affordable option. This follows expectations - greater privacy and more space typically command higher prices. Because the differences between room types are so distinct, this variable becomes an important categorical predictor for modeling.

City-level price differences are also visible. Cities like Hawaii, New York, and San Diego consistently show higher typical prices, while markets such as Columbus, Dallas, and Denver tend to have lower

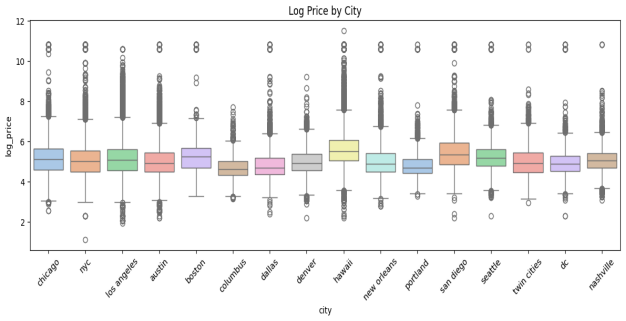

Figure 6. Log Price by City

show higher typical prices, while markets such as Columbus, Dallas, and Denver tend to have lower price levels. These patterns highlight that location is one of the strongest drivers of price variation. These differences also align with the earlier Listings per City pattern: cities with larger and more active Airbnb markets tend to show wider price ranges and more pronounced variation, likely because higher competition encourages hosts to differentiate their pricing strategies. This reinforces our thesis that supply concentration shapes price behavior across markets.

4. Listing Capacity vs Price

We also examined how property size relates to pricing using boxplots for bedrooms, bathrooms, and accommodates. All three follow similar upward trends: larger listings almost always have higher prices.These consistent patterns reinforce that listing size is one of the most important predictors of price

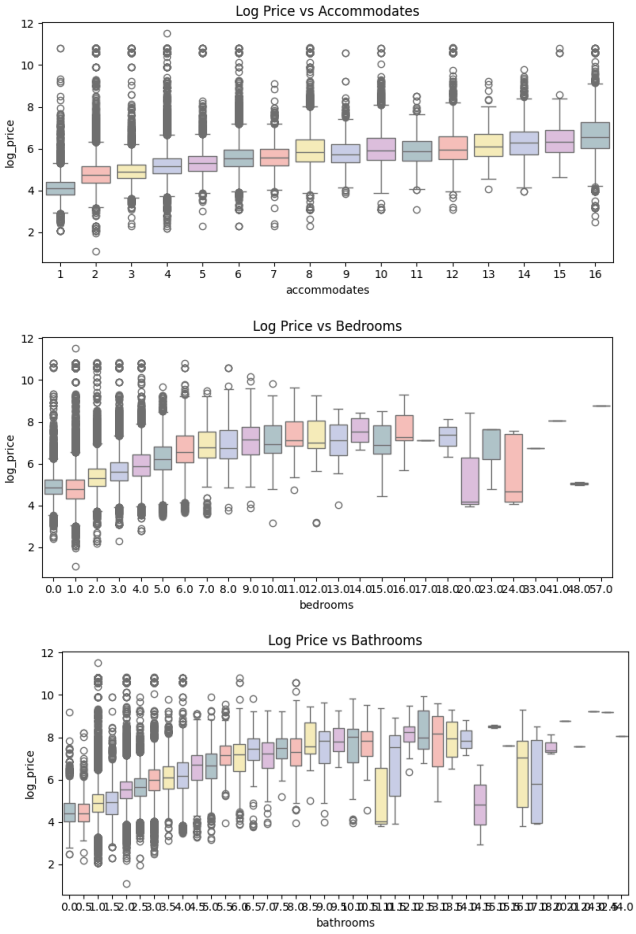

Figure 7. Listing Capacity vs Price

5. Reviews and Guest Demand Indicators

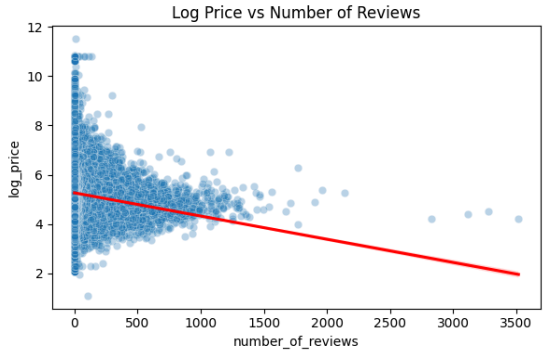

Figure 8. Log Price vs Number of Reviews

The scatter plot shows a weak negative trend: listings with many reviews often have lower prices. This likely happens because cheaper listings are booked more frequently and therefore receive more reviews. Meanwhile, high-priced listings may receive fewer stays and fewer total reviews.

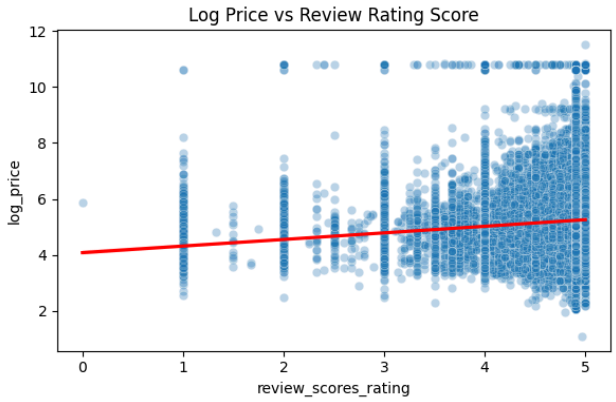

Figure 9. Log Price vs Review Rating Score

Review rating scores show a light positive association with price, but the effect is small. Most listings have ratings close to 5 stars, so the rating variable does not vary enough to strongly predict price. Even so, it still adds helpful information about guest satisfaction.

III. Methodology

A. Overview

To analyze the determinants of Airbnb listing prices and build predictive models, this study implements four regression approaches: Ordinary Least Squares (OLS), Ridge Regression, Lasso Regression, and Random Forest. These models were selected due to the large number of categorical variables in the dataset and the complex ways price varies across location, host behavior, property characteristics, and guest reviews. Each model followed a structured workflow involving data cleaning, standardization when appropriate, model fitting, and performance evaluation using an 80/20 train–test split.

Prior to estimation, the dataset was preprocessed to reduce sparsity in categorical indicators. In particular, neighborhood dummies were restricted to the five most represented neighborhoods within each city to avoid extreme imbalance. The dependent variable was defined as the natural logarithm of nightly price to address right-skewness. Independent variables included listing characteristics (accommodates, bedrooms, bathrooms, beds), host characteristics (response rate, acceptance rate, superhost status, identity verification), review scores (overall rating, accuracy, cleanliness, communication, location, and value), host tenure, activity measures, and full sets of categorical indicators for city, neighborhood, property type, and

room type. Because penalized models apply shrinkage based on coefficient magnitude, numerical variables were standardized using a StandardScaler to ensure comparability across predictors.

The four models complement one another. OLS establishes a baseline linear relationship, Ridge and Lasso address high dimensionality and multicollinearity among dummy variables, and Random Forest captures nonlinear patterns and interactions often present in housing markets.

B. Model 1: OLS Regression

The OLS model serves as the baseline for understanding how Airbnb listing characteristics relate to price under a standard linear specification. Using log-transformed price as the dependent variable, OLS estimates the marginal effect of each continuous feature while comparing average price differences across categorical factors such as neighborhood, room type, and property type. Because the dataset includes a large number of dummy variables—particularly for property type and neighborhood—OLS provides useful directional insights but may suffer from multicollinearity and unstable coefficient magnitudes. Despite these limitations, OLS offers a benchmark R^2 for comparison across models and helps identify which predictors appear most strongly associated with pricing before applying regularization techniques.

C. Model 2: Ridge Regression

For the second model, we extend the baseline specification by applying Ridge Regression to the same Airbnb dataset. We choose Ridge due to the structure of the data such as many listings share overlapping attributes, and the large number of city, neighborhood, room type, and property type dummies creates substantial multicollinearity. In this context, ordinary OLS coefficients can become unstable, particularly for categories with relatively few observations. Ridge Regression introduces an L2 penalty term that shrinks coefficient magnitudes toward zero without removing variables from the model. This allows us to retain the full richness of the dataset, including detailed neighborhood and property type indicators, while reducing overfitting and stabilizing estimates. Prior to fitting the Ridge model, all numerical predictors were standardized so that the penalty is applied on a comparable scale across variables. The model was estimated using RidgeCV, which selects the optimal regularization parameter α through five-fold cross-validation over a wide grid of candidate values.

D. Model 3: Lasso Regression

With similar reasons for choosing Ridge, we also extend our analysis with the third model, Lasso. It applies an L1 penalty to the same set of predictors in the Airbnb dataset to explore the role of feature selection alongside prediction. Lasso's L1 penalty differs from Ridge in that it can shrink some coefficients exactly to zero, effectively removing those predictors from the model. By zeroing out weaker predictors, it highlights which listing, host, and neighborhood characteristics are most strongly associated with nightly prices. Although this simplification can sometimes come at the cost of

slightly lower predictive performance compared to Ridge, it offers clear advantages in interpretability, helping to isolate the core drivers of Airbnb pricing in a complex, high-dimensional feature space. All numerical predictors were standardized, and LassoCV was used to select the optimal value of α via cross-validation.

E. Model 4: Random Forest

We implement Random Forest Regression as our final model to capture nonlinear relationships and interaction effects that linear and penalized regression models may miss. The Airbnb dataset includes rich combinations of attributes—for example, the way the effect of an extra bedroom might depend on neighborhood, property type, or review scores—which suggests that price formation may not be purely linear. Random Forest constructs an ensemble of decision trees using bootstrap samples of the Airbnb listings and random subsets of predictors at each split. Unlike the regression-based models, Random Forest does not require standardization of predictors and is relatively robust to multicollinearity and outliers in the dataset. We also use the same 80/20 train–test split to evaluate the model using test-sample R^2 and RMSE. Besides predictive performance, we see that feature importance measures derived from the fitted forest provide a complementary perspective on which variables such as accommodates, specific review scores, or particular neighborhoods play the largest role in explaining price variation. Although these importance metrics are less directly interpretable than regression coefficients, they offer valuable insight into the relative influence of different Airbnb attributes when nonlinearities and interactions are taken into account.

IV. Results

A. Model 1: OLS Regression

The Ordinary Least Squares (OLS) model achieve an R^2 of 0.607 (adjusted R^2 = 0.606), meaning the model explains about 61% of the variation in log nightly prices. The model is highly significant overall ($F = 582.9$, $p < 0.001$). However, the extremely small minimum eigenvalue (2.34e-26) indicates severe multicollinearity, largely due to the very large number of dummy variables. This justifies the need for regularized models like Ridge and Lasso to stabilize coefficients.

Listing attributes are among the strongest predictors. Bathrooms have the largest structural effect, with a coefficient of 0.1450, implying roughly a 15.6% price increase per additional bathroom. Bedrooms also increase price (0.0903), as does accommodates (0.0730), while beds show a negative association (-0.0197) once capacity variables overlap, demonstrating the multicollinearity among structural features. These results confirm that functional capacity, especially bathrooms, is the primary physical driver of price.

Review-based quality signals show selective influence. Cleanliness has a strong positive coefficient (0.1674), location also raises price (0.1330), and overall rating increases price (0.1551). In contrast, the value score is significantly negative

(–0.1604), and accuracy is statistically insignificant, proving that not all review dimensions influence pricing equally. The number of reviews has a very small negative effect (-0.0004), consistent with older or high-traffic listings tending to operate in more competitive or lower-priced markets.

Host behavior indicators are meaningful. Host response rate (0.0008) and host acceptance rate (0.0007) both significantly increase price, showing that responsive hosts can charge more. Surprisingly, Superhost status decreases price (–0.0229), suggesting many Superhosts specialize in affordable, high-occupancy units rather than premium listings—an interpretation supported directly by the negative coefficient.

Amenity effects reflect correlation structures rather than isolated preferences. Positive amenities include hair dryer (0.1269), carbon monoxide alarm (0.0784), and essentials (0.0633). Meanwhile, hot water (-0.2632), dishes/silverware (-0.2476), and kitchen (-0.0865) show negative coefficients, which demonstrates multicollinearity with property-type and neighborhood dummies rather than true negative value. The amenity count itself has a small positive effect (0.0062), confirming that individual amenity coefficients cannot be interpreted causally.

City-level differences reflect large regional price variation. Hawaii listings are significantly more expensive (+0.2254), as are NYC (+0.2019) and Boston (+0.1890). Conversely, Columbus (–0.7038), Dallas (–0.5520), Nashville (–0.2833), and New Orleans (–0.4223) are substantially cheaper. These values directly quantify how market conditions drive cross-city pricing.

Room type and property type produce some of the most dramatic differences. Hotel rooms have an extremely large positive coefficient (+2.4761), reflecting professionalized inventory, while private rooms (+0.2677) are priced below the reference entire-home baseline. Many unusual or low-frequency property types generate extreme coefficients, confirming again that multicollinearity and sparsity distort OLS estimates.

Neighborhood fixed effects remain among the strongest location drivers. Premium neighborhoods such as Lahaina (+0.3662), La Jolla (+0.3057), Belltown (+0.2978), and Midtown (+0.3624) raise prices dramatically, while areas such as Brighton (–0.3202), Dorchester (–0.3217), Long Beach (–0.2585), and Hennepin (–0.0743) correspond to lower-priced listings. These coefficients prove that hyper-local geography often outweighs city-level effects.

In summary, Airbnb pricing is primarily shaped by bathroom count (0.1450), guest capacity (0.0730), cleanliness (0.1674) and location ratings (0.1330), host responsiveness (0.0008–0.0007), and substantial city and neighborhood fixed effects. However, the extreme multicollinearity in the OLS model—evident from the eigenvalue (2.34e-26) and numerous contradictory signs—means many coefficients are unstable. Therefore, regularized models such as Ridge and Lasso are essential to

obtain reliable, interpretable pricing patterns beyond what OLS can provide.

B. Model 2: Ridge Regression

The Ridge regression model, tuned using 5-fold cross-validation, identifies $\lambda = 10$ as the optimal regularization strength. Using this penalty level, Ridge achieves an RMSE of 0.5850, MAE of 0.3891, and R^2 of 0.6026, performing nearly identically to OLS in explanatory power (OLS R^2 = 0.607) but producing far more stable and interpretable coefficients (Figure 10). This stability is visible in the coefficient path plot, where the top 15 influential predictors maintain smooth, regularized trajectories rather than the extreme swings observed in OLS due to multicollinearity.

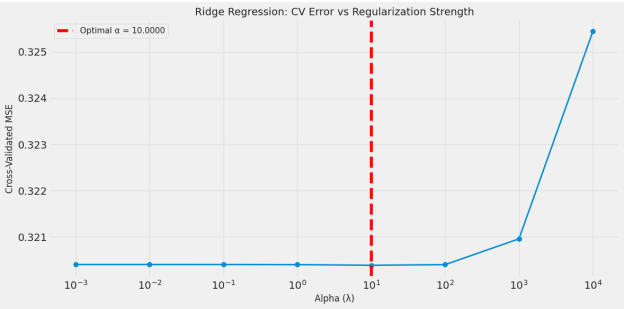

Figure 10. Ridge Regression CV Error vs Regularization Strength

In figure 11, it shows that Ridge highlights a similar set of core pricing drivers as OLS, but with cleaner, more coherent magnitudes. Bathrooms remain one of the strongest predictors, showing a coefficient of 0.13899, implying around a 14.9% price increase per additional bathroom—close to but slightly smaller than the OLS estimate. Accommodates (0.22523) and bedrooms (0.11379) also show large positive effects, reinforcing that functional capacity is a primary determinant of price. Amenities count contributes positively as well (0.10252), consistent with the idea that more equipped listings command higher rates, though Ridge shrinks individual amenity effects to avoid OLS-style exaggeration.

Location continues to exert major influence. Premium areas such as Lahaina (0.11171) and Midtown (0.06207) retain strong positive effects, while expensive product types—such as hotel rooms (0.25362) and entire serviced apartments (0.06653)—show some of the highest coefficients in the model. City-level signals persist too, with Hawaii (+0.07182) priced significantly above baseline. These results confirm that both micro- (neighborhood) and macro-level (city) geography remain central in Airbnb pricing.

Negative coefficients are notably more reasonable under Ridge, reflecting the shrinkage effect. Cities like Dallas (–0.08599) and Columbus (–0.09553) continue to show price discounts, consistent with their lower tourism demand. Property types such as private rooms in rental units (–0.23547), private rooms in homes (–0.18755), and entire rental units (–0.16417) remain systematically cheaper than the baseline category. Amenity effects also stabilize: features like hot water (–0.10083) and dishes/silverware (–0.09888) no longer carry the extreme and unrealistic magnitudes observed in OLS, illustrating Ridge's ability to suppress noise created by collinearity among amenity and property-type features.

Compared to OLS, Ridge produces substantially more stable coefficients, avoids the inflated or contradictory signs caused by multicollinearity, and eliminates the extreme swings across dummy variables. While OLS had severe numerical instability (minimum eigenvalue 2.34e-26) and many implausibly large categorical effects, Ridge shrinks these into realistic ranges while preserving the underlying structure of the relationships. Ridge's predictive accuracy ($R^2 = 0.6026$) is nearly identical to OLS ($R^2 = 0.607$), but its interpretation is far more reliable, showing that regularization is essential for this high-dimensional, dummy-heavy Airbnb pricing model.

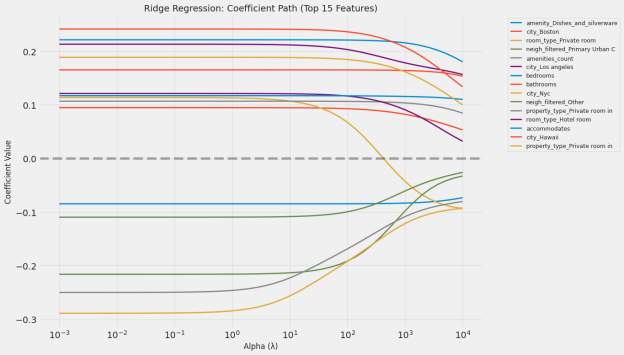

Figure 11. Ridge Regression: Coefficient Path (top 15 Features)

C. Model 3: Lasso Regression

Lasso regression was utilized to identify the most influential predictors of Airbnb pricing while performing feature selection through L1 regularization. Cross-validation identified an optimal penalty level of $\alpha = 0.001$, resulting in an R^2 of approximately 0.609 and an RMSE of 0.568. The cross-validation plot (Figure 12) shows the pattern where larger λ values overshrink coefficients and degrade predictive accuracy. The minimum of the CV curve at $\alpha = 0.001$ reflects the best trade-off between bias and variance, producing the most stable and generalizable solution. At this value, the model retains 199 predictors while shrinking 34 predictors to exactly zero, illustrating the sparsity-inducing nature of L1 regularization.

The Lasso coefficient path plot (Figure 13) provides insight into the regularization process. As λ increases, less important predictors shrink toward zero rapidly, while the most influential variables

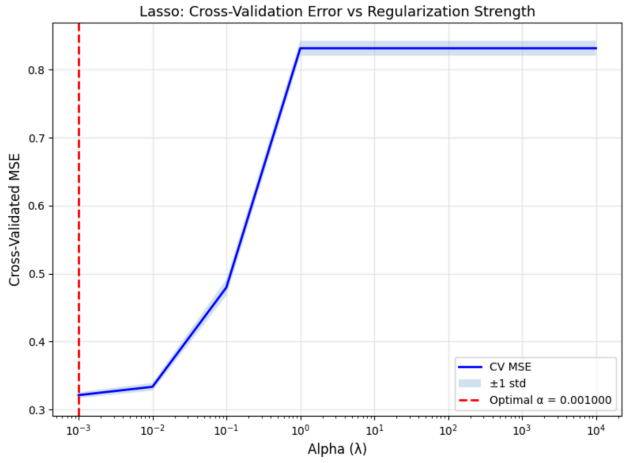

Figure 12. Lasso Cross-validation plot

maintain substantial magnitude until very close to the optimal α. This behavior reflects how Lasso effectively identifies the strongest structural drivers of price while discarding weaker or redundant information. Among the retained coefficients, several emerged as major contributors.

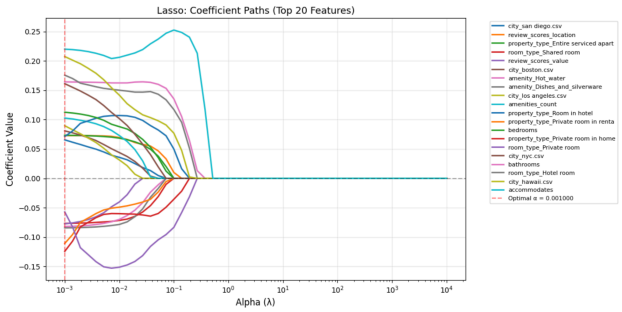

Figure 13. Lasso Coefficient Path Plot

On the positive side, accommodation is the strongest predictor (β = 0.220), reaffirming that listing capacity is a primary determinant of price. Several city indicators also have large positive coefficients, including Hawaii, New York City, Los Angeles, and Boston, reflecting the substantial price premiums associated with high-demand markets. Structural characteristics such as bathrooms, bedrooms, and amenities_count also exhibit strong positive coefficients, suggesting that both property size and amenity richness significantly elevate willingness to pay. The strong weight on room_type: Hotel room indicates that hotel-style units operate at systematically higher price points compared to residential private rooms or shared accommodations.

The negative coefficients are equally informative. Lasso assigns large negative values to private-room formats, such as Private room in home and Private room in rental unit, confirming that this segment occupies the lower end of the price distribution relative to entire-unit rentals or hotel-style listings. Neighborhood indicators, such as Primary Urban Center and Other (within particular cities), also contribute negatively, suggesting that certain neighborhoods systematically depress prices regardless of property characteristics. Interestingly, certain amenities, such as Hot water and Dishes_and_silverware, receive negative coefficients. These amenities are extremely common across listings; their negative coefficients likely reflect their association with lower-end or older units rather than causal downward pressure on price. This illustrates how Lasso captures correlations in the data structure rather than interpreting amenities in isolation.

Overall, the Lasso model provides a highly interpretable, sparse representation of price determinants across a large and heterogeneous dataset. Its feature selection reveals clear patterns: prices are driven upward by listing size, amenity richness, and location in high-value markets, while lower-cost segments, such as private rooms, peripheral neighborhoods, and low-end property types, are associated with downward pricing pressure. Importantly, Lasso achieves these insights while maintaining strong predictive performance, demonstrating the value of regularization when working with high-dimensional, multicollinear, and highly categorical Airbnb data.

D. Model 4: Random Forest

The Random Forest model applied to the Airbnb dataset demonstrates strong predictive performance and provides valuable insights into the drivers of Airbnb listing prices. Overall, the model achieves a test R^2 of 0.756, indicating that it explains approximately 75.6% of the variance in log-transformed prices on unseen data. This

performance is complemented by a test RMSE of 0.449, suggesting that predictions deviate from the true log(price) by roughly 0.45 units on average.

Despite its strong performance, the model displays signs of overfitting, such as the 0.208 gap between training and test R^2 values. The model still generalizes well to new data, but the degree of overfitting highlights that the current hyperparameters allow individual trees to grow too complex. The extensive set of over 200 features created from dummy encoding further contributes to this tendency, giving the model space to capture spurious patterns. The nature of Airbnb Pricing Data, which highly fluctuates based on seasonality, events, and demand, also plays an essential role. Nevertheless, the model still exhibits high fidelity on the training set and strong generalization on the test set, providing a solid foundation for further analysis.

Predicted versus actual values (Figure 14) reinforce these observations. In the training set, predictions fall tightly along the 45-degree perfect prediction line, reflecting the model's near-perfect fit. Small vertical bands are visible at high price levels (log(price) > 9). Lower price regions exhibit minimal variance, again confirming the model's high accuracy on familiar data. In contrast, the test set displays noticeably wider dispersion around the diagonal line. Predictions for mid-priced listings (log(price) between 4 and 8) remain well-calibrated, but substantial underprediction occurs for luxury listings, and mild overprediction appears for low-priced properties. This suggests that the model is highly reliable for average Airbnb listings, typically in the $50–$300 range, but caution is warranted when pricing high-end properties, where unusual features or location factors not captured in the dataset play a larger role.

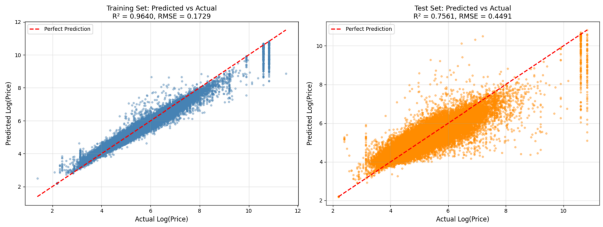

Figure 14. Predicted vs. Actual Plots

A central strength of the Random Forest approach lies in its feature importance analysis (Figure 15), which reveals interpretable insights into price determinants. The top five predictors, accommodates, bedrooms, beds, bathrooms, and amenities_count, collectively capture a large portion of model influence, aligning closely with real-world pricing logic. Accommodation capacity alone accounts for 8.56% of total importance, while bedroom and bathroom counts contribute an additional 6-7% each. Amenities also play a significant role, signaling that well-equipped listings command higher prices. While city and neighborhood variables appear later in the rankings, their contributions, though smaller, remain meaningful, reflecting geographic effects. Host attributes and review scores show lower individual importance but contribute collectively to model accuracy, supporting the idea that property characteristics matter most in setting prices, while host reputation and location reinforce but rarely dominate.

Compared to linear models such as Ridge or Lasso regression, Random Forest produces superior predictive accuracy and richer insights into nonlinear

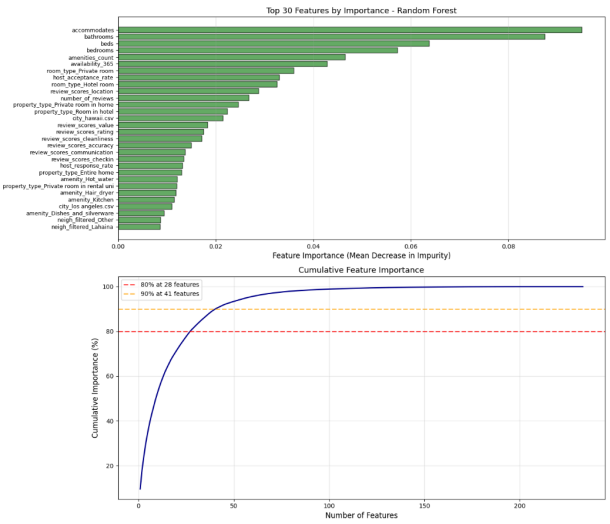

Figure 15. Feature Importance Analysis

interactions. The model performs strongly in predicting Airbnb prices and provides clear evidence that accommodates, property size, and amenities are the dominant drivers of price variation across listings. Although the model displays moderate overfitting and reduced accuracy for luxury properties, its ability to capture nonlinear structure and complex interactions makes it an effective tool for pricing analysis.

V. Discussion & Recommendations

A. FIFA 2026

Although the models in this study capture Airbnb pricing dynamics under typical market conditions, the upcoming 2026 FIFA World Cup introduces an important external shock that may temporarily reshape the determinants identified in our analysis. According to Deloitte's economic impact report, the World Cup is expected to draw nearly five million visitors to the sixteen host cities—Atlanta, Boston, Dallas, Houston, Kansas City, Los Angeles, Miami, New York–New Jersey, Philadelphia, San Francisco–Bay Area, Seattle, and others—resulting in significant short-term accommodation pressures. Airbnb prices during the event are projected to rise by roughly 90 percent compared to comparable summer periods, approximately 2.7 million nights booked between June 11 and July 19, 2026.

One of the most direct implications is the anticipated increase in nightly accommodation prices. The estimated 90% increase mirrors historical patterns documented in major sporting events such as Super Bowls or Formula One races, where hotel prices have surged between 70% and 180%. Realtor.com similarly notes that past World Cups have triggered dramatic price spikes in host cities, with hotels and short-term rentals reaching historically high occupancy rates as the event draws near.

Beyond price effects, the World Cup is poised to significantly reshape short-term rental supply capacity. Airbnb's official partnership with FIFA highlights the platform's role in expanding accommodation options across dispersed neighborhoods, particularly where hotel infrastructure is insufficient. Deloitte reports show that Airbnb listings reach 67% of ZIP codes in U.S. host cities, compared to only 38% coverage by hotels. This spatial elasticity is crucial that unlike hotels, the Airbnb supply can expand by activating existing housing stock. Survey evidence from host

cities indicates that 68%–84% of residents would consider hosting during the World Cup, depending on the city. This surge in supply not only increases the number of available listings but shapes the geographic distribution of tourism spending, which pushes economic activity into residential neighborhoods that do not traditionally benefit from major events.

Another major macroeconomic effect of the World Cup is the expected increase in guest spending, which directly influences housing demand. Deloitte estimates that Airbnb guests will generate approximately 2.7 million guest nights during the World Cup and over USD 1.2 billion in direct spending across North America. Average visitor spending is forecasted at USD 500+ per night, with about USD 150–180 allocated to accommodation depending on the city. As accommodation spending makes up a significant share of total tourist expenditure, fluctuation in listing prices becomes an important determinant of how travelers distribute their budgets across cities and neighborhoods. In this sense, Airbnb pricing is not merely a market phenomenon but an integral part of the broader economic ripple effect—driving indirect and induced impacts on retail, transit, food, and entertainment sectors.

Future research can build on this analysis by incorporating actual Airbnb data from the 2026 World Cup once the event occurs. With repeated observations over time, researchers could construct a panel dataset for the sixteen host cities and comparable non-host cities, enabling the use of causal inference methods. A difference-in-differences (DiD) design could compare price trends before and after the event in host versus non-host cities to isolate the causal impact of the World Cup on both nightly price levels and supply responses. Also, future work could explore how machine-learning models adapt under extreme conditions by incorporating event indicators or city-by-event interaction terms into forecasting models such as Random Forests or gradient-boosted trees. These techniques would allow researchers to examine shifts in variable importance during mega-events and assess whether nonlinear models outperform linear ones when markets face exceptional demand pressures.

B. Market Structure - Supply & Demand Across Cities

To understand how market structure differs across Airbnb cities, we first examine the composition of supply using property-type and room-type distributions. The bar charts (Figure 16, 17) reveal that “Entire place” listings overwhelmingly dominate supply across every cities, especially visible in Hawaii and Los Angeles. This pattern is consistent with tourism-oriented areas where travelers prioritize privacy and standalone units.

A second notable pattern is the presence of substantial private-room segments in high-density, high-rent cities such as New York City, Los Angeles,

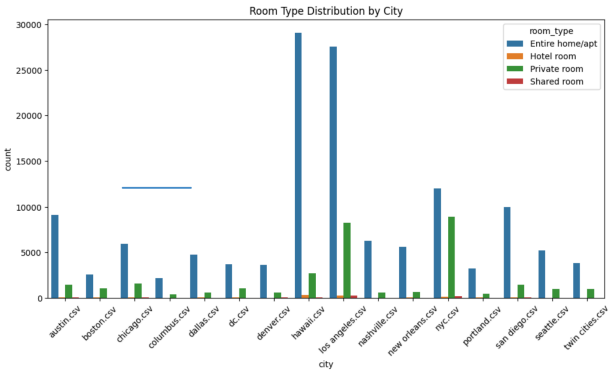

Figure 16. Room Type Distribution by City

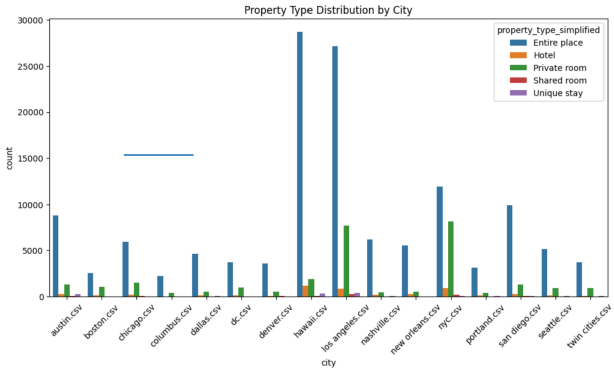

Figure 17. Property Type Distribution by City

and Chicago. New York City alone contains close to 8,000 private rooms, making it the largest private-room market in the dataset. These private-room listings cater to budget travelers, students, temporary workers, and other demand groups seeking lower-cost alternative accommodations. In contrast, smaller or more suburban markets, such as Columbus, Portland, Denver, and Nashville, show low private-room supply, indicating that their Airbnb ecosystems are more oriented toward full-unit stays.

Using three variables, total supply, reviews per month, and availability, we derived a supply-demand imbalance metric that approximates market slack (oversupply) or tightness (undersupply). The results show that Los Angeles (309.16), Hawaii (241.69), and NYC (182.58) exhibit the largest imbalance ratios by wide margins. These values are driven by enormous listing counts (ranging from 21,000 to 36,000) combined with lower-than-average demand intensity, measured through reviews per month. The result is a large supply base chasing relatively slower-moving demand. These imbalance values imply that hosts in these cities face some of the highest levels of competition in the entire dataset, which may lead to more aggressive price differentiation and undercutting, findings later corroborated by the price dispersion analysis. Notably, these markets are also those with the strongest tourism identities, yet their supply expansions appear to have outpaced demand growth.

The wide range in imbalance ratios (from ~20 to >300) highlights the heterogeneity of Airbnb market conditions. High-imbalance markets (LA, Hawaii, NYC) may face intensified price competition and greater price dispersion, while low-imbalance markets (Columbus, Portland) have tighter supply conditions that help sustain more stable pricing.

The price-dispersion statistics confirm and deepen these structural insights. Cities exhibit substantial variation in average prices, standard deviations, and inequality (P90/P10 ratios). Cities such as Hawaii (mean = $906, std = 4519) and San Diego (mean = $706, std = 4023) display exceptionally high distributions and large tails. These cities also show extremely high P90/P10 ratios (≈ 7 - 8), indicating that the top 10% of listings are priced 7 - 8 times higher than the bottom 10%. These markets reflect

the presence of luxury properties, unique stays, and vacation rentals that strongly skew the upper end of the price distribution. The enormous standard deviations (>4000 for Hawaii) further support the presence of megahigh-priced outliers.

Price dispersion is strongly correlated with supply imbalance. Oversupplied markets (LA, Hawaii, NYC) also show the widest dispersion, reflecting both competition-driven discounting and the existence of ultra-luxury listings. Meanwhile, tight markets (Denver, Columbus) display more compact price distributions, consistent with more predictable demand and limited extreme-tier supply.

Across all analyses, a coherent pattern emerges: Airbnb markets differ substantially in supply composition, competitive intensity, and pricing structure, with these dimensions tightly interrelated. Oversupplied metropolitan markets like Los Angeles, Hawaii, and NYC show extremely high supply-demand imbalance scores, large concentrations of entire-home listings, and some of the most unequal price distributions. In contrast, smaller or more regionally focused markets such as Columbus, Portland, and Denver exhibit tighter supply–demand conditions, more uniform supply structures, and narrower price spreads. The regression analysis reinforces the idea that both supply scale and booking intensity jointly shape market-level pricing, reflecting the interaction between host behavior and underlying demand. Overall, the extension demonstrates that Airbnb pricing cannot be understood solely through micro-level listing characteristics - market structure and city-level dynamics play equally critical roles in shaping price outcomes.

VI. Conclusion

This study provides a comprehensive analysis of Airbnb pricing across 16 major cities in the United States. The results show that property capacity, number of bathrooms and bedrooms, and available amenities are the most influential listing-level determinants of nightly price. Geographic factors, particularly at the neighborhood level, also contribute strongly to price differences. Among the predictive approaches, the Random Forest model performs the best, which suggests that price formation in short term rental markets involves nonlinear relationships that are better captured by machine learning methods.

The extended analysis of market conditions demonstrates that pricing outcomes are shaped by more than only individual property attributes. Cities with large supply expansion tend to experience stronger price competition and greater price dispersion, while markets with more limited availability maintain steadier pricing. This finding highlights the importance of understanding how supply and demand interact at the city level when evaluating pricing strategies and market performance.

The discussion of the upcoming 2026 FIFA World Cup illustrates how major tourism events can temporarily influence both supply availability and pricing behavior in short term rental markets. These external demand surges represent a valuable

opportunity to study how market dynamics shift under unusual conditions and to test whether predictive models remain reliable during periods of rapid change.

Although the models used in this study offer strong predictive accuracy, future research should incorporate data from multiple time periods in order to evaluate changes over time. Methods that rely on event indicators, causal inference, and policy variables may also provide greater insight into how short term rental markets respond to regulatory actions and large tourism events. Overall, the findings show that Airbnb pricing is driven by a combination of property characteristics, broader competitive conditions, and external tourism pressures, and that continued data-driven analysis will be essential for understanding the evolution of short term rental markets.

References

Chen, C.-F., & Rothschild, R. (2010). An application of hedonic pricing analysis to the case of hotel rooms in Taipei. Tourism Economics, 16(3), 685–694. https://doi.org/10.5367/000000010792278310

Carvell, Steven A. and Herrin, William E. (1990) "Pricing in the Hospitality Industry: An Implicit Markets Approach," Hospitality Review: Vol. 8: Iss. 2, Article 3. https://digitalcommons.fiu.edu/hospitalityreview/vol8/iss2/3

Teubner, T., Hawlitschek, F., & Dann, D. (2017). Price determinants on Airbnb: How reputation pays off in the sharing economy. Journal of Self-Governance & Management Economics, 5(4), 53–80. https://publikationen.bibliothek.kit.edu/1000068696

Zhang, Z., Ye, Q., & Law, C. H. R. (2011). Determinants of hotel room price: An exploration of travelers' hierarchy of accommodation needs. International Journal of Contemporary Hospitality Management, 23(7), 972-981. https://doi.org/10.1108/09596111111167551

Jiao, J., & Bai, S. (2020). An empirical analysis of Airbnb listings in forty American cities. Cities, 99, 102618. https://doi.org/10.1016/j.cities.2020.102618

Guttentag, D. A., & Smith, S. L. (2017). Assessing Airbnb as a disruptive innovation relative to hotels: Substitution and comparative performance expectations. International Journal of Hospitality Management, 64, 1–10. https://scholarworks.umass.edu/items/821ce6fd-259b-45b3-aa86-e03869cac805

Deboosere, R., Kerrigan, D. J., Wachsmuth, D., & El-Geneidy, A. (2019). Location, location and professionalization: a multilevel hedonic analysis of Airbnb listing prices and revenue. Regional Studies, Regional Science, 6(1), 143–156. https://doi.org/10.1080/21681376.2019.1592699

Barron, K., Kung, E., & Proserpio, D. (2017). The sharing economy and housing affordability.

https://www.aeaweb.org/conference/2018/preliminary/paper/ykYrh4Gd

Horn, K., & Merante, M. (2017). Is home sharing driving up rents? Evidence from Airbnb in Boston. Journal of Housing Economics, 38, 14–24. https://doi.org/10.1016/j.jhe.2017.08.002

Wachsmuth, D., Kerrigan, D., Chaney, D., & Shillolo, A. (2017). Short-term cities: Airbnb's impact on Canadian housing markets. Policy report. Urban Politics and Governance research group, School of Urban Planning, McGill University.

Ert, E., Fleischer, A., & Magen, N. (2016). Trust and reputation in the sharing economy: The role of personal photos in Airbnb. Tourism Management, 55, 62–73. https://www.sciencedirect.com/science/article/pii/S0261517716300127

Lee, D., Hyun, W., Ryu, J., Lee, W. J., Rhee, W., & Suh, B. (2015). An analysis of social features associated with room sales of Airbnb. Paper presented at the ACM Conference Companion on Computer Supported Cooperative Work & Social Computing, Vancouver, British Columbia. https://doi.org/10.1145/2685553.2699011

Almeida, A., Nunes, A. P., & Machado, L. P. (2025). How Do Reviews Impact Airbnb's Prices? A Hedonic Approach. Tourism and Hospitality, 6(4), 181. https://doi.org/10.3390/tourhosp6040181

Lin, W., & Yang, F. (2024). The price of short-term housing: A study of Airbnb on 26 regions in the United States. Journal of Housing Economics, 65, 102005. https://doi.org/10.1016/j.jhe.2024.102005

Kalehbasti, H., Singh, A., & Rao, H. R. (2021). Airbnb price prediction using machine learning and sentiment analysis. International Cross-Domain Conference for Machine Learning and Knowledge Extraction, pp. 173-184. Springer, Cham, 2021 https://doi.org/10.1007/978-3-030-84060-0_11

Chapman, S., Mohammad, S., & Villegas, K. (2023). Predicting Listing Prices In Dynamic Short Term Rental Markets Using Machine Learning Models. arXiv preprint arXiv:2308.06929.

Medpalliwar, A., Choube, D., Kute, G., Kaushik, K., & Meshra, P. (2025). Airbnb rental price prediction using machine learning techniques. In AIP Conference Proceedings (Vol. 3233, Issue 1, Article 020039). https://doi.org/10.1063/5.0245572

Alharbi, Z. H. (2023). A Sustainable Price Prediction Model for Airbnb Listings Using Machine Learning and Sentiment Analysis. Sustainability, 15(17), 13159. https://doi.org/10.3390/su151713159

Camatti, N., di Tollo, G., Filograsso, G. et al. Predicting Airbnb pricing: a comparative analysis of artificial intelligence and traditional approaches. Comput Manag Sci 21, 30 (2024). https://doi.org/10.1007/s10287-024-00511-4

Deloitte. (2025). The economic impact of Airbnb during the FIFA World Cup 2026. Airbnb. https://news.airbnb.com/wp-content/uploads/sites/4/

2025/06/Deloitte-Report_Airbnb_FIFAWorldCup26.pdf

Airbnb. (2024, May 15). Airbnb and FIFA announce major multi-tournament partnership. https://news.airbnb.com/airbnb-and-fifa-announce-major-multi-tournament-partnership/

Realtor.com. (2024, June 24). Will Airbnb prices surge during the 2026 FIFA World Cup? Here's what to expect. https://www.realtor.com/news/trends/fifa-world-cup-airbnb/