Airbnb Pricing Analysis Report

Dillon Welindt

2025-09-18

Contents

1 Executive Summary

This project aims to test two pricing models for AirBnB listings using data from airroi.com. I examined 6,272 listings within 1,500 properties across five cities (New York City, Chicago, Los Angeles, Houston). I compared a bayesian regression-based model against XGBoost. Unsurprisingly, XGBoost was a more accurate pricing tool, though the inferences from the bayesian model were maintained.

Results indicate that location within city, number of bedrooms, and number of baths are the most important predictors of price. Seasonality is not a strong predictor. There is substantial heterogeneity in occupancy within listings, likely based on the host's appetite for business, as the link between price and occupancy is not as strong as anticipated.

Topographical pricing maps are presented and can be used in conjunction with real estate listings to estimate a likely ROI on a property.

Qualitative analysis demonstrates that amenities serve primarily as an indicator of listing type. One notable exception to this is that amenities related to childcare (high chair, toys, etc.) command a premium.

2 Introduction

2.1 Background

AirBnB has fundamentally changed the short-term rental market by challenging traditional hotel industry dynamics. This market transformation has created both opportunities and challenges for property investors and hosts. Short-term rental success depends on complex interactions between property characteristics, location attributes, seasonal demand patterns, and local market dynamics. Hosts must navigate pricing decisions that balance occupancy rates with revenue maximization.

The complexity of Airbnb pricing optimization stems from several factors. First, the market exhibits strong spatial heterogeneity - properties within the same city can command vastly different rates based on neighborhood characteristics, proximity to attractions, and local amenities. Second,

temporal demand patterns may vary significantly across markets, with some cities showing strong seasonal variation while others maintain relatively stable year-round demand. Third, property characteristics interact in non-linear ways - a swimming pool may command a substantial premium in Miami but provide minimal value in New York City.

Host decision-making is frequently guided by platform recommendations that use basic algorithmic approaches or manual adjustments based on limited market visibility. This creates opportunities for more sophisticated analytical approaches that can identify the true drivers of pricing power and revenue optimization.

3 Data and Methods

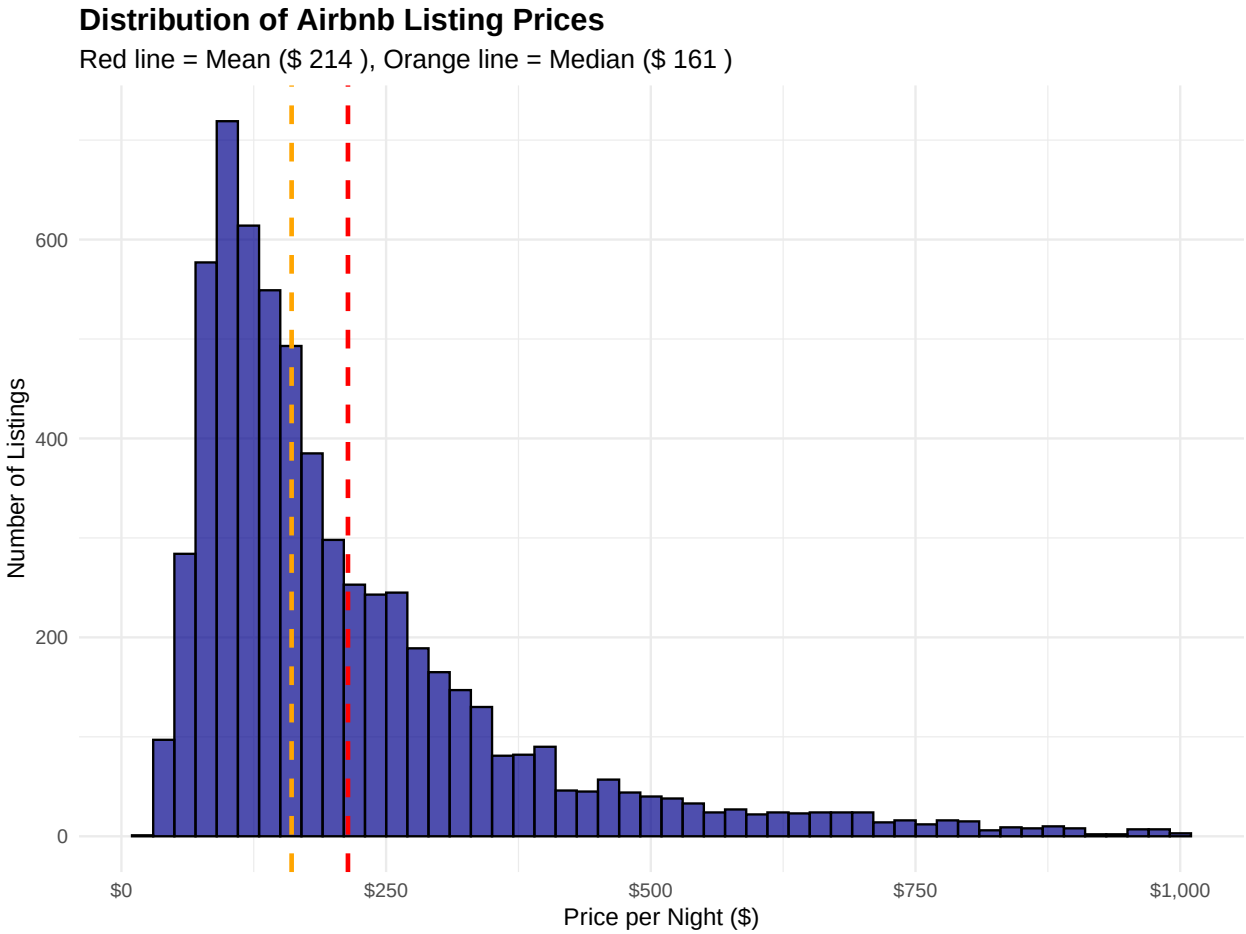
3.1 Data Description

The dataset comprises publicly available data from airroi.com, covering five major US metropolitan areas: New York City, Chicago, Los Angeles, Houston, and Miami. The data include both property characteristics (location, bedrooms, bathrooms, amenities, property type) and monthly pricing/occupancy aggregates. Data are separated by city and by data type: property or pricing. The former include information on the unit. The latter are monthly aggregates per property of bookings. The data are from August 1, 2024 through August 1, 2025. Airroi limits these data to 300 properties per city, and the selection process is unknown.

Table 1: Dataset Summary Statistics

Metric	Value
Total Listings	6,272
Total Properties	984
Cities	5
Date Range	Monthly aggregates
Price Range	\$30 - \$1,000 (M= 213.9, SD=159.1)
Average Occupancy	58.3%

These summary statistics inadequately describe prices, which were approximately exponentially distributed (see figure below). Prices primarily range from \$100-250 per night (10th percentile was 79.7, 90th was 411.9).



As noted, these data are joined from property and listings (monthly usage per property). We can examine the distribution of listings across properties (table below). These summary statistics demonstrate that approximately 80% of listings are from half of the total properties. Conversely, approximately 22% of properties are listed only once or twice. Given that the data provided by airroi are a subset of their entire data (limited to 300 properties per city), it is unclear whether this is due to sampling, or a true representation of the AirBnB market. In the case of the latter, this would indicate turnover in market supply–that is, AirBnB units have either been introduced or removed from the platform within the calendar year.

Table 2: Property Duplication Breakdown

Category	# Properties	Total Listings	% Properties	% Listings
1 listing	117	117	11.9	1.9
2 listings	107	214	10.9	3.4
3 listings	88	264	8.9	4.2
4-5 listings	154	694	15.7	11.1
6-10 listings	284	2256	28.9	36.0
Eleven+ listings	234	2727	23.8	43.5

3.2 Methodology Overview

Data features extracted include month, season, and hypothesized busy periods (March, December-January, and March). Price was log-transformed. Spatial clusters were also derived using k-means clustering of listings based on their geographic location (latitude/longitude).

Listings costing more than $1000/night were removed as luxury units likely follow different trends than standard rentals. Units with less than 10% occupancy were also removed. In the data exploration process, it was noted that occupancy was significantly zero-inflated. This likely reflects a decision on the part of the host to not rent and thus their price as a salient predictor of occupancy is moot. Thus, the dataset represents properties with active occupancy, focusing on operational listings rather than inactive inventory.

The geographic locations of the properties in this set are shown below split by city and colored by spatial cluster.

3.2.1 Hierarchical Bayesian Model

The Bayesian models employ a hierarchical structure using the brms package to capture nested geographical and temporal relationships in Airbnb pricing. The approach models log-transformed prices as a function of standardized property characteristics (beds, bathrooms), listing type, and seasonal indicators, while incorporating multiple levels of random effects.The hierarchical structure includes spatial clustering (properties nested within geographic clusters) and city-season interactions, allowing the model to account for both local neighborhood effects and city-specific seasonal patterns. Priors based on intuition - for example, expecting positive effects for bedrooms/bathrooms and modest seasonal variations. Domain expertise would be valuable, although the data should be sufficient to overwhelm mis-specified priors. The model uses Hamiltonian Monte Carlo sampling (4 chains, 2000 iterations) to generate posterior distributions, providing uncertainty quantification for all parameter estimates.

This approach excels at interpretability, offering clear coefficient estimates with interpretable posterior distributions, and handles the nested spatial structure naturally through its hierarchical framework.

Four models were tested. All models use a hierarchical structure in which units are situated within neighborhoods in cities and command different prices. Additionally, the latter models have a temporal parameter for either month or season, again with differential effects by city.

```
Model 1:
log_price ~ log_price ~ beds_std + baths_std +
    (1 | city) + (1 | spatial_cluster)

Model 2:
log_price ~ beds_std + baths_std + season
    + is_weekend + (1 | season:city) + (1 | spatial_cluster)

Model 3:
log_price ~
    beds_std + baths_std + listing_type+season + is_weekend +
    (1 | spatial_cluster) + (1 | month:city)
```

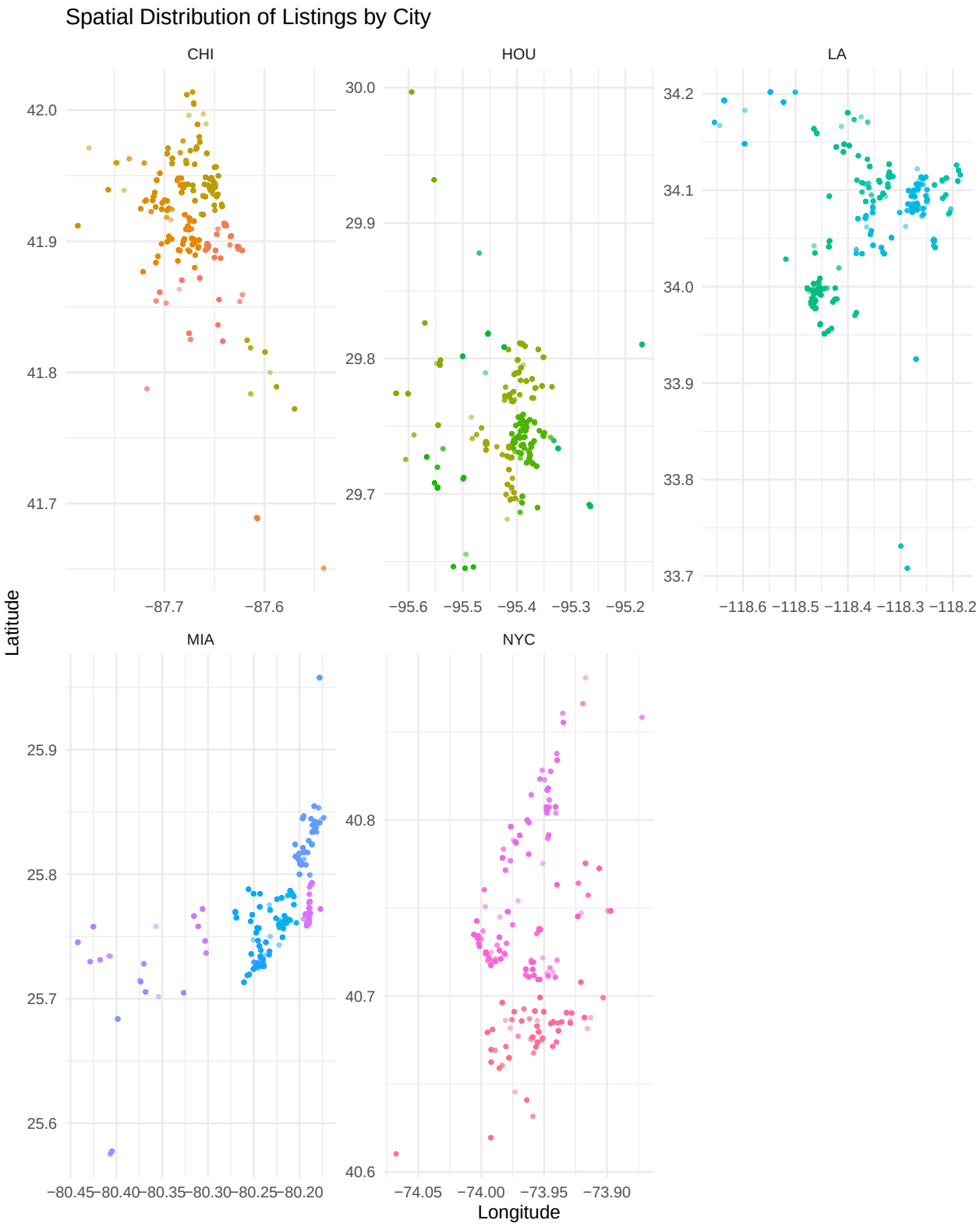


Figure 1: Spatial Distribution of Listings by City

```
Model 4:
log_price ~
    beds_std + baths_std + listing_type + season +
    (1 | spatial_cluster) + (1 | month:city)
```

Based on the expected log pointwise predictive density estimated from cross-validation, Model 4 was ultimately selected. This model is simpler than Model 3, as it omits the is_weekend variable, which was minimally impactful.

3.2.2 XGBoost Machine Learning

The XGBoost implementation uses gradient boosting to capture complex non-linear relationships and feature interactions in the pricing data. Categorical variables (listing type, season, city) were transformed into one-hot encoded features and created interaction terms (particularly season-city combinations) to capture location-specific temporal patterns. The approach automatically learns feature interactions and non-linear relationships without requiring explicit specification.

The model architecture uses 500 boosting rounds with early stopping, a learning rate of 0.1, maximum tree depth of 6, and regularization through subsampling (80% of data and features per tree). This configuration balances model complexity with generalization capability.

XGBoost provides feature importance rankings based on gain, cover, and frequency metrics, revealing which variables contribute most to prediction accuracy. The model optimizes a regression objective on log-transformed prices. This approach prioritizes predictive accuracy over interpretability (complementing the Bayesian approach), excelling at capturing complex patterns in the data that linear models might miss.

3.2.3 Amenities Text Analysis

The amenities analysis employs a text processing pipeline to extract value-driving features from unstructured (free-text) amenity descriptions. Given the imbalance among total number of listings per property, amenities are examined at the property level. I parse comma-separated amenities into a binary feature matrix, identifying single amenities with at least 100 occurences in these data.

For each amenity, it computes the average price among listings that offer it versus the overall market baseline. Price premiums are calculated both in absolute terms and as percentage differences using log-price transformations to account for the multiplicative nature of pricing effects.

4 Results

4.1 Bayesian Model Results

4.1.1 Model Performance

The hierarchical Bayesian model demonstrates strong predictive performance with excellent convergence diagnostics (all Rhat = 1.00). The model achieved a log-scale correlation of 0.80 with actual prices, indicating good fit for the multiplicative pricing relationships inherent in real estate markets.

4.1.2 Key Findings from Bayesian Analysis

Table 3: Random Effects Variance Components

	Grouping Factor	Parameter	Std Dev	Variance
city	city	Intercept	0.270	0.073
spatial_cluster	spatial_cluster	Intercept	0.368	0.135
residual___	residual___		0.406	0.165

4.1.3 Key Findings from Bayesian Analysis

Property Characteristics Effects:

- **Bedrooms**: Each additional bedroom increases price by 21%
- **Bathrooms**: Each additional bathroom increases price by 20%

These effects are nearly equivalent, contradicting conventional wisdom that bedrooms are more valuable than bathrooms. The tight confidence intervals (0.18-0.20 for beds, 0.16-0.21 for baths) indicate robust estimates.

Listing Type Hierarchy: The model reveals a clear hierarchy of property types, with entire homes commanding large premiums over shared accommodations:

- **Entire Serviced Apartment**: +110% premium
- **Entire Home**: +73% premium
- **Entire Loft**: +70% premium
- **Entire Townhouse**: +60% premium

Conversely, hostel accommodations show significant discounts:

- **Private Room in Hostel**: -54% discount
- **Room in Hotel**: -53% discount

Temporal Patterns: Seasonal effects are modest but statistically sound (compared to winter):

- **Summer**: +7% premium
- **Spring**: +3% premium
- **Fall**: Minimal effect (-1%)

The limited seasonal variation suggests that spatial and property-type factors dominate temporal considerations in pricing.

Spatial Effects: The hierarchical structure reveals important spatial patterns:

- **Spatial Clusters**: phi = 0.37 (substantial neighborhood-level variation)
- **City-level**: phi = 0.27 (moderate city-level differences)

The large spatial cluster variation indicates that location within a city matters significantly, with some neighborhoods commanding substantial premiums over others.

4.2 XGBoost Machine Learning Results

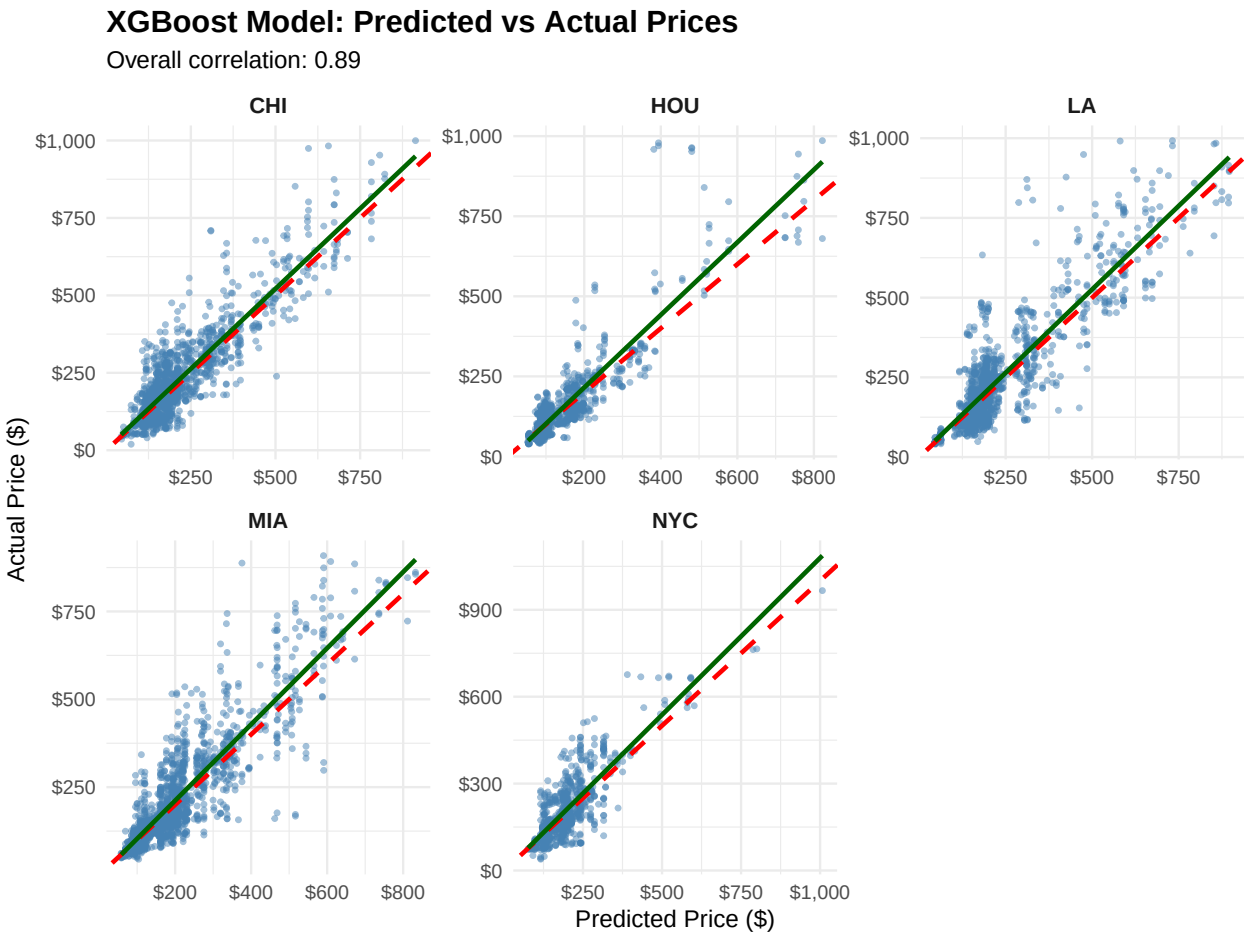
4.2.1 Model Performance and Comparison

The XGBoost model demonstrates superior predictive performance compared to the Bayesian approach, particularly for complex non-linear relationships. Performance metrics across training, test, and Bayesian model comparisons reveal:

Table 4: Model Performance Comparison

Model	Log_Correlation	Price_Correlation	Log_RMSE	Price_RMSE	MAPE
XGBoost Train	0.878	0.895	0.306	72.73	24.09
XGBoost Test	0.861	0.870	0.328	78.67	26.02
Bayesian	0.796	0.694	0.390	115.55	31.79

4.2.2 XGBoost Diagnostics



Key Performance Insights:

1. **Superior Accuracy**: XGBoost achieves 86-87% correlation on test data vs 79% for Bayesian
2. **Better Generalization**: Small train-test gap (0.878 vs 0.861) indicates good generalization
3. **Lower Prediction Errors**: MAPE of 26% vs 32% for Bayesian model
4. **Robust Performance**: Consistent performance across log and price scales

4.2.3 Feature Importance Analysis

The XGBoost model reveals the relative importance of different features in predicting Airbnb prices. Unlike traditional regression coefficients, these importance scores capture non-linear relationships and feature interactions.

Feature Importance Hierarchy:

The machine learning approach confirms and extends the Bayesian findings:

1. **Bathrooms Dominate (47% of model gain)**: Bathroom count is the single most important predictor, substantially outweighing all other features
2. **Spatial Location (30% of model gain)**: Spatial cluster membership remains crucial for pricing
3. **Bedrooms Secondary (19% of model gain)**: Bedroom count is important but clearly secondary to bathrooms
4. **Seasonal Effects Minimal (<2% combined)**: Temporal patterns provide minimal predictive value

Table 5: Top 10 XGBoost Feature Importance Metrics

Feature	Gain	Cover	Frequency
baths_std	0.467	0.140	0.202
spatial_cluster_num	0.301	0.384	0.319
beds_std	0.192	0.150	0.259
season_summer	0.005	0.024	0.033
summer_CHI	0.005	0.026	0.013
season_fall	0.004	0.021	0.034
season_spring	0.004	0.022	0.032
fall_MIA	0.003	0.022	0.011
summer_LA	0.003	0.024	0.012
spring_LA	0.003	0.029	0.010

4.2.4 Key Machine Learning Insights

Non-Linear Relationships: The superior performance of XGBoost suggests important non-linear relationships in the data that the linear Bayesian model cannot capture. This is particularly evident in:

- **Bathroom-price relationships**: May exhibit threshold effects or diminishing returns

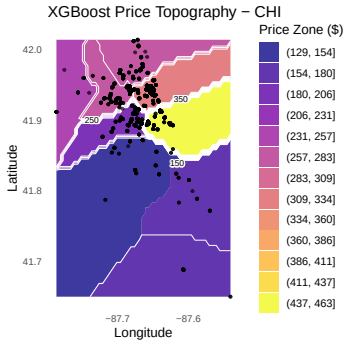
- **Spatial clustering**: Complex geographic patterns beyond linear distance effects
- **Property type interactions**: Combinations of features that create premium segments

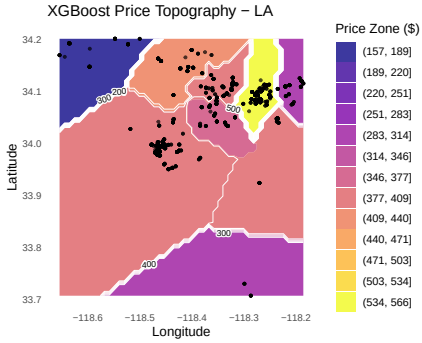
Feature Interactions: XGBoost automatically captures interactions between features, revealing that pricing is not simply additive across characteristics. The model learns complex patterns such as:

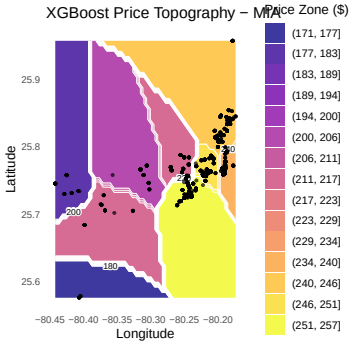
- High-bathroom properties in premium locations commanding disproportionate premiums
- Seasonal effects varying by location and property type
- Property characteristic interactions that create market segments

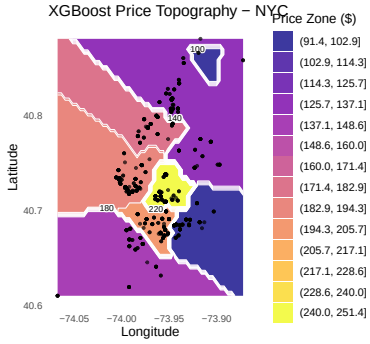
Temporal Patterns Confirmed: Both models agree that seasonal effects are minimal, with XGBoost allocating less than 2% of its predictive power to temporal features. This validates the Bayesian finding that spatial and property characteristics dominate pricing.

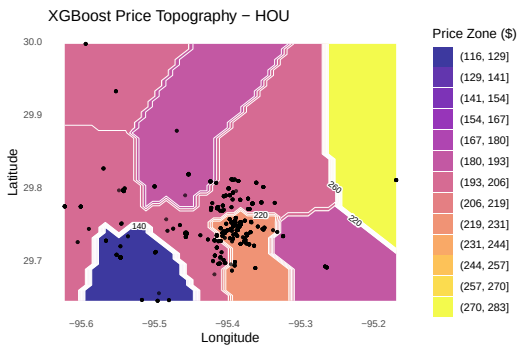
Spatial Price Topography The XGBoost model enables the creation of price topographical maps that reveal spatial pricing patterns across metropolitan markets.


XGBoost Price Topography − CHI
Price Zone ($)
(129, 154]
(154, 180]
(180, 206]
(206, 231]
(231, 257]
(257, 283]
(283, 309]
(309, 334]
(334, 360]
(360, 386]
(386, 411]
(411, 437]
(437, 463]
Latitude
Longitude
42.0
41.9
41.8
41.7
−87.7
−87.6
350
250
150


XGBoost Price Topography − LA
Price Zone ($)
(157, 189]
(189, 220]
(220, 251]
(251, 283]
(283, 314]
(314, 346]
(346, 377]
(377, 409]
(409, 440]
(440, 471]
(471, 503]
(503, 534]
(534, 566]
Latitude
Longitude
34.2
34.1
34.0
33.9
33.8
33.7
−118.6
−118.5
−118.4
−118.3
−118.2
300
200
500
300
400


XGBoost Price Topography − MIA
Price Zone ($)
(171, 177]
(177, 183]
(183, 189]
(189, 194]
(194, 200]
(200, 206]
(206, 211]
(211, 217]
(217, 223]
(223, 229]
(229, 234]
(234, 240]
(240, 246]
(246, 251]
(251, 257]
Latitude
Longitude
25.9
25.8
25.7
25.6
−80.45
−80.40
−80.35
−80.30
−80.25
−80.20
200
180


XGBoost Price Topography − NYC
Price Zone ($)
(91.4, 102.9]
(102.9, 114.3]
(114.3, 125.7]
(125.7, 137.1]
(137.1, 148.6]
(148.6, 160.0]
(160.0, 171.4]
(171.4, 182.9]
(182.9, 194.3]
(194.3, 205.7]
(205.7, 217.1]
(217.1, 228.6]
(228.6, 240.0]
(240.0, 251.4]
Latitude
Longitude
40.8
40.7
40.6
−74.05
−74.00
−73.95
−73.90
100
140
180
220


XGBoost Price Topography − HOU
Price Zone ($)
(116, 129]
(129, 141]
(141, 154]
(154, 167]
(167, 180]
(180, 193]
(193, 206]
(206, 219]
(219, 231]
(231, 244]
(244, 257]
(257, 270]
(270, 283]
Latitude
Longitude
30.0
29.9
29.8
29.7
−95.6
−95.5
−95.4
−95.3
−95.2
140
220
260
220

4.2.5 Business Implications from Machine Learning

The XGBoost results provide several actionable insights:

1. **Bathroom Priority**: The 47% importance score for bathrooms suggests this is the primary lever for pricing optimization
2. **Location Optimization**: 30% importance for spatial clusters indicates significant revenue opportunities through location-based pricing
3. **Simplified Seasonal Strategy**: Minimal temporal importance suggests complex seasonal pricing may not justify the operational complexity
4. **Property Type Focus**: The model's ability to capture non-linear relationships suggests property-specific pricing strategies may be more effective than broad-based approaches

4.3 Amenities Analysis

4.3.1 Most Common Amenities

The most common amenity (n=981) listed was wifi. The following were also present in at least 90% of properties: smoke alarm, "essentials", heating, air conditioning, hot water, iron, tv, hangers, kitchen, hair dryer, dishes and silverware, refrigerator. Notably, washers and dryers were quite common (n=778 and n=771, respectively).

4.3.2 Price Premium Analysis

In this section, I compare the average price of units containing various amenities against the average of price of those without. Averages are compared within unit types (e.g., entire home, apartment, single-room), and weighted by number of occurences. The difference between the prices is the premium represented by that amenity. Note that amenities may be an indicator of unit type or other aspect beyond the amenity. For instance, "laundromat near by" also implies lack of washer/dryer. See table below.

Table 6: Top 15 Amenities by Weighted Price Premium

Amenity	Properties	Weighted Premium %
high chair	132	41.21
indoor fireplace	166	40.34
sound system	101	35.09
pack 'n play/travel crib	250	32.06
gym	133	28.70
children's books and toys	149	25.65
elevator	167	25.01
crib	141	24.11
hot tub	134	24.03
board games	111	20.79
paid parking off premises	115	20.64

dishwasher	471	20.26
barbecue utensils	147	19.01
bbq grill	293	13.64
cleaning before checkout	121	12.00

The most notable amenities with minimal property type spillover are those associated with barbecuing appliances and childcare. Both have a strong premium associated. However, there are additional risks associated with each. Barbecuing may add considerable fire risk and would increase costs, assuming the host needs to supply gas periodically. Children may also be more likely to cause damage to the unit and increase cleaning costs. Amenities associated with each (though especially childcare) should be further explored for their probably return-on-investment.

4.4 Price-Occupancy Relationships

4.4.1 Occupancy Modeling

A separate Beta regression model was fitted for occupancy rates:

```
occupancy ~ log_price + beds_std + baths_std + season + listing_type +
            (1 + log_price | city:season) + (1 | spatial_cluster)
```

Key Findings: Baseline Occupancy is 84.8%.

Price Effect: -0.38 (95% CI: -0.47 to -0.29) log-price coefficient This negative coefficient confirms higher prices reduce occupancy rates (i.e., a 10% price increase corresponds to roughly a 3-4 percentage point decrease in occupancy)

Property Characteristics: Occupancy rates were generally consistent among the most common property types, although it should be noted that private rooms in a shared environment were rare in these data. Estimates can be provided for occupancy, assuming baseline levels for other parameters. Common types and their effects are:

- **Entire home:** .70 - Higher occupancy than baseline. Occupancy 91.8% at baseline.
- **Private room in townhouse:** .48 - Strong positive effect. Occupancy 90.0% at baseline.
- **Entire rental unit:** .61 - Strong occupancy performance. Occupancy 91.1% at baseline.
- **Entire guest suite:** .56 - Strong occupancy rates. Occupancy 90.7% at baseline.

On the low end, occupancy rate for a private room in a hostel was 69.2%, though this was a small sample.

All seasonal effects are modest.

4.4.2 Pricing Demand Curves by City

One of the most surprising findings from this analysis is the relatively weak negative relationship between price and occupancy rates. While the occupancy model confirms that higher prices reduce occupancy (coefficient = -0.38, $p < 0.001$), the overall correlations between price and occupancy

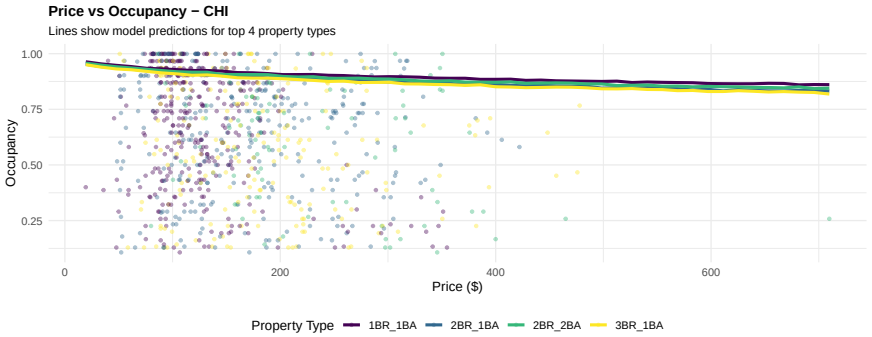
across cities range from only 0.09 to 0.22. This modest relationship contradicts traditional economic theory, which would predict a stronger inverse relationship between price and demand.

Several factors likely explain this counterintuitive finding:

Market Efficiency Through Competitive Pricing: The weak price-occupancy relationship may indicate that most hosts have already optimized their pricing strategies relative to their local markets. If properties are generally priced competitively within their respective market segments, we would expect to observe exactly this pattern - prices that reflect quality and location premiums while maintaining similar occupancy rates across different price points.

Market Segmentation Effects: Different price points may serve distinct customer segments with varying demand elasticities. Business travelers and luxury vacation renters may be relatively price-insensitive, while budget-conscious leisure travelers demonstrate higher price sensitivity. This segmentation can obscure overall price-demand relationships in aggregate analyses.

This finding has important implications for revenue optimization strategies. Rather than focusing primarily on price adjustments to drive occupancy, hosts may achieve better results by emphasizing property differentiation, strategic positioning within their quality tier, and understanding their target market segment's specific preferences and price sensitivity.


Price vs Occupancy – CHI
Lines show model predictions for top 4 property types
Occupancy
1.00
0.75
0.50
0.25
0
200
400
600
Price ($)
Property Type
1BR_1BA
2BR_1BA
2BR_2BA
3BR_1BA


Price vs Occupancy – LA
Lines show model predictions for top 4 property types
Occupancy
1.00
0.75
0.50
0.25
200
400
600
Price ($)
Property Type
1BR_1BA
2BR_1BA
2BR_2BA
3BR_1BA


Price vs Occupancy – MIA
Lines show model predictions for top 4 property types
Occupancy
1.00
0.75
0.50
0.25
100
200
300
400
500
Price ($)
Property Type
1BR_1BA
2BR_1BA
2BR_2BA
3BR_1BA


Price vs Occupancy – NYC
Lines show model predictions for top 4 property types
Occupancy
1.00
0.75
0.50
0.25
200
400
600
Price ($)
Property Type
1BR_1BA
2BR_1BA
2BR_2BA
3BR_1BA


Price vs Occupancy – HOU
Lines show model predictions for top 4 property types
Occupancy
1.00
0.75
0.50
0.25
100
200
300
Price ($)
Property Type
1BR_1BA
2BR_1BA
2BR_2BA
3BR_1BA

5 Discussion

5.1 Business Implications

There are two clear revenue optimization opportunities: the first is to identify undervalued real estate for AirBnBs. The second is to position current AirBnBs to families with young children.

Although different areas command different prices, the overall range of prices may be more constricted than the property values of these units. Put plainly, properties in less-desirable parts of a city may still be valuable as AirBnBs. This may be especially true of areas with effective public transit or paid transportation services (e.g., Uber, Lyft)–which did not appear in the amenities analysis.

Purchasing childcare equipment and marketing the unit as child-friendly would likely have a positive ROI. This would be dependent on unit characteristics; units with multiple bedrooms and appropriate storage may benefit most from this.

Further analysis should done to identify demand curves for price points within units.

5.2 Limitations

These data are limited in scope, as they include only 300 properties per city. This would be especially concerning if there is a systematic selection bias in determining their publication. The amenities analysis is exploratory and could be incorporated in a causal inference framework.

6 Conclusions

The findings here are largely intuitive: cities have clusters of neighborhoods which command different rates. Seasonal effects more limited than anticipated: these cities may consistently attract tenants such that demand stays high, seasonality may be more limited and require additional granularity, or hosts may not tend to adjust prices to meet demand. This latter finding seems partially true, given the heterogeneity in occupancy rates. Last, and perhaps most importantly, there is evidence that–within market range–pricing units higher does not reduce demand sufficient to hurt revenue.

6.1 Future Research

Further analysis could be done to develop within-unit patterns. A bayesian approach to assessing the impact of childcare or barbecue equipment would allow for a simulated counterfactual to test my marketing conjecture. This work could be extended beyond these five cities by simply downloading additional data and adding it within the loading section. This would improve the stability of the Bayesian posteriors. The incremental value of public transportation could be tested either by identifying proximity to such stops (difficult) or by comparing cities with stronger public transportation than others. Additional work should also consider beds/baths as indicators of listing types.