



UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS

DÉPARTEMENT DES SCIENCES ADMINISTRATIVES

WORKING PAPER No. 4

Hedonic Pricing of Short-Term Rental Properties: A Multi-Method Analysis of Airbnb Listing Prices in Quebec

Simon-Pierre Boucher

Département des sciences administratives

Université du Québec en Outaouais

simon-pierre.boucher@uqo.ca

Gatineau – Pavillon Alexandre-Taché

283, boulevard Alexandre-Taché

Gatineau, Québec, Canada J9A 1L8

May 2026

Version 1.0

Abstract

This paper develops a comprehensive hedonic pricing model for short-term rental properties listed on Airbnb across Quebec, Canada. Drawing on [Rosen \(1974\)](#) and [Lancaster \(1966\)](#), we decompose nightly prices into implicit valuations of structural characteristics, amenity bundles, host attributes, reputation signals, locational factors, and text-derived features. Our dataset comprises 5,000 listings from Inside Airbnb, of which 3,412 retain valid prices after cleaning. We employ a layered econometric strategy: nine nested OLS specifications with HC3 robust standard errors, progressing from a structural-only model ($R^2 = 0.052$) to a full specification with spatial controls ($R^2 = 0.274$); quantile regression at $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$ revealing heterogeneity across the price distribution; machine learning methods—XGBoost ($R^2 = 0.285$), Random Forest, Gradient Boosting, LightGBM—with SHAP interpretability; and spatial econometric models—spatial lag, spatial error, and spatial Durbin specifications—accounting for spatial dependence. Bedrooms, room type, review scores, spatial context, and amenity indices are the dominant price determinants, with their relative importance shifting across quantiles and geographic sub-markets.

Keywords: hedonic pricing, Airbnb, short-term rentals, quantile regression, spatial econometrics, machine learning, SHAP, Quebec, amenity valuation

JEL Classification: C21, C31, C45, C58, L83, R31, R33

1. Introduction

The theory of hedonic prices, formalized by [Rosen \(1974\)](#) and rooted in the characteristics approach to consumer demand developed by [Lancaster \(1966\)](#), posits that differentiated goods can be described by a vector of objectively measured attributes and that the market price of such a good equals the sum of the implicit valuations consumers place on each attribute. In the four decades since Rosen's seminal contribution, hedonic models have become a workhorse of applied microeconomics, with extensive applications in real estate ([Chattopadhyay and Machemehl, 2006](#)), automobiles, environmental valuation, and, more recently, the sharing economy.

The rapid global expansion of Airbnb and competing short-term rental (STR) platforms has created a uniquely rich laboratory for hedonic analysis. Unlike traditional hotel room inventories, which are relatively standardized within a star-rating tier, Airbnb listings exhibit enormous heterogeneity along structural, locational, amenity, and reputational dimensions. Each listing is, in effect, a bundle of characteristics chosen and priced by an individual host operating in a decentralized marketplace. The platform's publicly accessible data—including detailed descriptions, amenity lists, review scores, host profiles, calendar availability, and geographic coordinates—provides researchers with an unusually granular view of the hedonic price surface.

A growing empirical literature has exploited this richness. [Wang and Nicolau \(2017\)](#) estimate hedonic price equations for a cross-section of Airbnb listings in 33 major cities, finding that property type, location, host characteristics, and online reputation are significant price determinants. [Gibbs et al. \(2018\)](#) extend the analysis to include amenity bundles and find that premium amenities such as pools and hot tubs command substantial price premiums. [Zhang et al. \(2017\)](#) and [Chen and Xie \(2020\)](#) document similar patterns across Chinese and U.S. markets, respectively, while [Benítez-Aurioles \(2018\)](#) and [Dogru and Pekin \(2017\)](#) examine European and Turkish markets. [Pérez-Sánchez et al. \(2018\)](#) apply geographically weighted regression to Spanish Airbnb data, demonstrating that implicit prices vary spatially. [Gunter and Önder \(2018\)](#) study Vienna's market structure, while [Magno et al. \(2018\)](#) investigate the role of host experience in shaping pricing outcomes.

Despite these contributions, several gaps remain. First, most studies rely on a single econometric specification—typically ordinary least squares (OLS) applied to a log-linear price equation—and therefore cannot capture heterogeneity across the conditional price distribution. Second, the potential for complex non-linear interactions among listing characteristics is seldom explored, and when machine learning methods are employed, the resulting models are treated as black boxes rather than as complements to structural hedonic inference. Third, spatial dependence in listing prices—a first-order concern in any market where location matters—is frequently acknowledged but rarely modeled explicitly using spatial econometric techniques.

This paper addresses all three gaps by developing a comprehensive multi-method hedonic pricing model for Airbnb listings in Quebec, Canada. Our central research questions are:

1. What are the implicit prices of structural, amenity, host, reputation, locational, and text-derived characteristics of Airbnb listings in Quebec?
2. How do these implicit prices vary across the conditional price distribution, from budget to luxury segments?
3. To what extent do non-linear interactions and spatial dependence alter the hedonic price surface estimated by conventional linear models?
4. Which listing attributes contribute most to predictive accuracy, and does the importance ranking differ between econometric and machine learning frameworks?

Our contribution is fourfold. First, we assemble a comprehensive hedonic dataset for Quebec—a market that, despite its significance as a major North American tourism destination, has received relatively little attention in the Airbnb pricing literature. Our sample encompasses 5,000 listings across Montréal, Quebec City, the Laurentians, the Eastern Townships, Charlevoix, and dozens of smaller municipalities. Second, we construct a feature space spanning seven attribute categories—structural characteristics, amenity indices, host attributes, reputation signals, locational variables, NLP-derived text features, and spatial controls—producing 49 modeling features. Third, we implement a layered estimation strategy combining OLS with heteroskedasticity-robust standard errors (HC3), quantile regression at five percentiles, an ensemble of seven machine learning models interpreted through SHAP values (Lundberg and Lee, 2017), and spatial econometric models including spatial lag, spatial error, and spatial Durbin specifications (LeSage and Pace, 2009; Anselin, 1988). Fourth, we demonstrate how estimated implicit prices aggregate into a price decomposition that benchmarks any listing against the market, providing actionable intelligence for hosts and platform managers.

The remainder of the paper is organized as follows. Section 2 describes the data sources, variable construction, and summary statistics. Section 3 presents the econometric and machine learning methodology. Section 4 reports the main empirical results. Section 5 discusses robustness checks. Section 6 concludes.

2. Data

2.1. Data Source

The primary data for this study are drawn from Inside Airbnb (<http://insideairbnb.com>), an independent, open-source project that periodically scrapes publicly available listing information

from the Airbnb platform. The extracts provide a snapshot of every active listing, including the nightly price, property and room type, geographic coordinates, host profile variables, review scores, amenity counts, and free-text listing names and descriptions.

Our sample covers the province of Quebec, Canada, encompassing the major urban markets of Montréal, Quebec City, Gatineau, Sherbrooke, and Trois-Rivières, as well as popular tourism and resort regions including the Laurentians (Mont-Tremblant, Saint-Sauveur), the Eastern Townships (Magog, Bromont), the Charlevoix corridor, the Saguenay–Lac-Saint-Jean region, and the Gaspésie coast.

2.2. Sample Construction

Starting from the raw extract of 5,000 listings with 19 original variables, we apply the following cleaning procedures:

1. **Price validity.** Listings with a nightly price below \$10 CAD (17 observations) or above \$25,000 CAD (27 observations) are set to missing, eliminating data-entry errors and placeholder prices. After this step, 3,412 listings retain valid price observations.
2. **Price transformations.** We construct the raw price, three Winsorized versions at the 1st/99th percentiles (\$38–\$10,822), 2nd/98th (\$46–\$8,117), and 5th/95th (\$63–\$4,921), plus natural log transformations.
3. **Rating standardization.** Scores originally on a 0–100 scale are converted to 0–5. Of the 5,000 listings, 810 (16.2%) lack a rating.
4. **Geocoding validation.** Listings whose coordinates fall outside the geographic boundaries of Quebec (latitude $\in [44.5, 55.5]$, longitude $\in [-80.0, -56.0]$) are removed (6 observations).
5. **Property type standardization.** Approximately 30 Inside Airbnb property type variants are mapped to six canonical categories: Rental unit, House, Cabin/Chalet, Condo, Apartment, and Other.

After applying these filters, the analytical sample comprises $N = 3,412$ listings with valid prices. Models incorporating additional covariates with their own missing values use smaller effective samples (reported alongside each specification).

2.3. Variable Definitions

We organize the variables into seven categories. The dependent variable is the natural logarithm of the nightly listing price in Canadian dollars ($\ln p_i$). The log transformation addresses the strong right skew

of the raw price distribution (mean = \$1,115; median = \$219; standard deviation = \$2,253) and yields coefficients interpretable as approximate percentage changes via $\text{Premium}(\%) = 100 \times (\exp(\hat{\beta}) - 1)$.

Structural characteristics ($\mathbf{S}_i$) capture the physical attributes: bedrooms (mean = 2.08), bathrooms (mean = 1.28), guest capacity (mean = 5.31), log-transformed counterparts, per-guest ratios, property type dummies (entire home = 48.1%, private room = 35.5%), a large-capacity indicator (≥ 8 guests), quadratic and interaction terms.

Amenity indices ($\mathbf{A}_i$) are constructed from the raw amenity count (mean = 44.9, range 3–109): luxury index, business travel index, family comfort index, leisure index, remote work index, and amenity richness index, all normalized to $[0, 1]$.

Host attributes ($\mathbf{H}_i$) include Superhost status (52.3%), Guest Favorite badge (33.4%), multi-listing host indicator, and a Superhost $\times$ entire-home interaction.

Reputation variables ($\mathbf{R}_i$) comprise the overall review score (mean = 4.80 on a 0–5 scale), log review count, quality score, reputation score, cleanliness premium indicator, and missing-rating flag.

Locational variables ($\mathbf{L}_i$) include Haversine distance to the nearest city centre (22 municipalities), local listing density within 500m (mean = 14.7, max = 125) and 1 km, local average price within 500m and 1 km, and a K -means spatial cluster ($K = 20$).

Text-derived features ($\mathbf{N}_i$) include title length, seven bilingual keyword scores, and 10 TF-IDF/SVD embedding components capturing 16.98% of lexical variance.

2.4. Descriptive Statistics

Table 1 reports descriptive statistics for the key variables. The price distribution is strongly right-skewed: the mean (\$1,115) exceeds the median (\$219) by a factor of five, confirming the appropriateness of the log transformation.

Table 2 presents the price distribution by property type. Condos and apartments exhibit the highest mean prices, driven by urban premium listings, while Cabin/Chalet listings show moderate means with substantial variation reflecting Quebec's diverse cottage and ski resort market.

2.5. Correlation Structure

Table 3 presents pairwise Pearson correlations among the key continuous variables. Bedrooms, bathrooms, and guest capacity are moderately to strongly intercorrelated ($r = 0.64$ – 0.81), motivating the inclusion of per-guest ratios and log transformations to mitigate multicollinearity. The number of reviews shows a negative correlation with log price ($r = -0.24$), suggesting that more frequently booked listings tend to be lower-priced.

Table 1. Summary Statistics of Key Variables

Variable	Mean	Std. Dev.	Min	P25	Median	P75	N
Price (\$CAD)	1,115	2,253	10.5	125.0	219.2	1,139	3,412
Log price	5.87	1.42	2.36	4.83	5.39	7.04	3,412
Bedrooms	2.08	1.53	0	1	2	3	5,000
Bathrooms	1.28	1.05	0	1	1	2	5,000
Guest capacity	5.31	3.37	1	2	4	7	5,000
Amenities count	44.9	16.0	3	34	45	56	5,000
Rating (0–5)	4.80	0.23	2.25	4.73	4.86	4.96	4,190
Num. reviews	62.7	85.0	0	10	33	82	4,275
Quality score	93.8	41.9	0	94.5	107.9	121.0	5,000
Num. images	27.9	19.2	0	14	23	36	5,000
Is Superhost	0.52	0.50	0	0	1	1	5,000
Is Guest Favorite	0.33	0.47	0	0	0	1	5,000
Pets allowed	0.29	0.46	0	0	0	1	5,000

Notes: This table reports descriptive statistics for the main variables used in the empirical analysis. Price is measured in Canadian dollars per night. Rating is on a 0–5 scale. Binary variables are coded as 0/1.

Table 2. Price Distribution by Property Type (\$CAD)

Property Type	Mean	Median	Std. Dev.	Min	Max	N
Apartment	1,276	201	3,328	15	18,117	77
Cabin/Chalet	969	237	2,215	48	23,321	691
Condo	1,344	210	2,343	49	14,577	237
House	1,058	223	2,206	18	22,811	679
Other	1,203	170	2,727	11	23,625	554
Rental unit	1,136	247	1,925	14	24,928	1,174

Notes: This table reports the distribution of nightly prices (CAD) by property type for listings with valid prices ($N = 3,412$).

Table 3. Correlation Matrix of Key Variables

	log_price	bed	bath	guests	amen.	rating	reviews
log_price	1.000						
Bedrooms	0.111***	1.000					
Bathrooms	0.138***	0.704***	1.000				
Guests	0.072***	0.806***	0.637***	1.000			
Amenities	−0.051**	0.322***	0.290***	0.370***	1.000		
Rating	0.006	0.129***	0.157***	0.113***	0.274***	1.000	
Reviews	−0.236***	−0.061**	−0.049**	−0.029	0.095***	0.072***	1.000

Notes: Pearson correlation coefficients. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

2.6. Price Distribution

Figure 1 displays the distribution of nightly prices. The raw price distribution is strongly right-skewed, with the median (\$219) substantially below the mean (\$1,115). Figure 2 shows the log-transformed distribution, which is approximately symmetric.

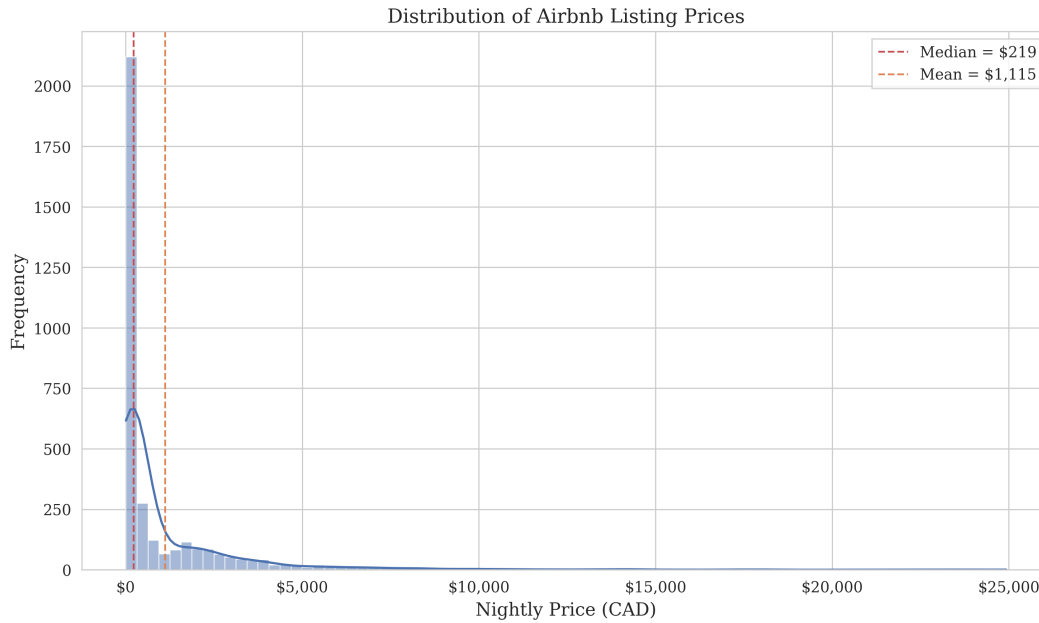


Figure 1. Distribution of Nightly Listing Prices (\$CAD)

Notes: This figure plots the histogram and kernel density estimate of nightly listing prices in Canadian dollars. The vertical dashed lines indicate the mean and median.

2.7. Geographic Coverage

Figure 3 maps the spatial distribution of listings across Quebec, with colour indicating nightly price. The majority of listings are concentrated in Montréal, with secondary clusters in Quebec City, the Laurentian resort corridor, and the Eastern Townships. Local listing density reaches 125 within 500 metres in central Montréal.

2.8. Price Variation by Property Type and Neighbourhood

Figure 4 displays boxplots of nightly prices for the top 20 cities by listing count, and Figure 5 shows the price distribution by property type.

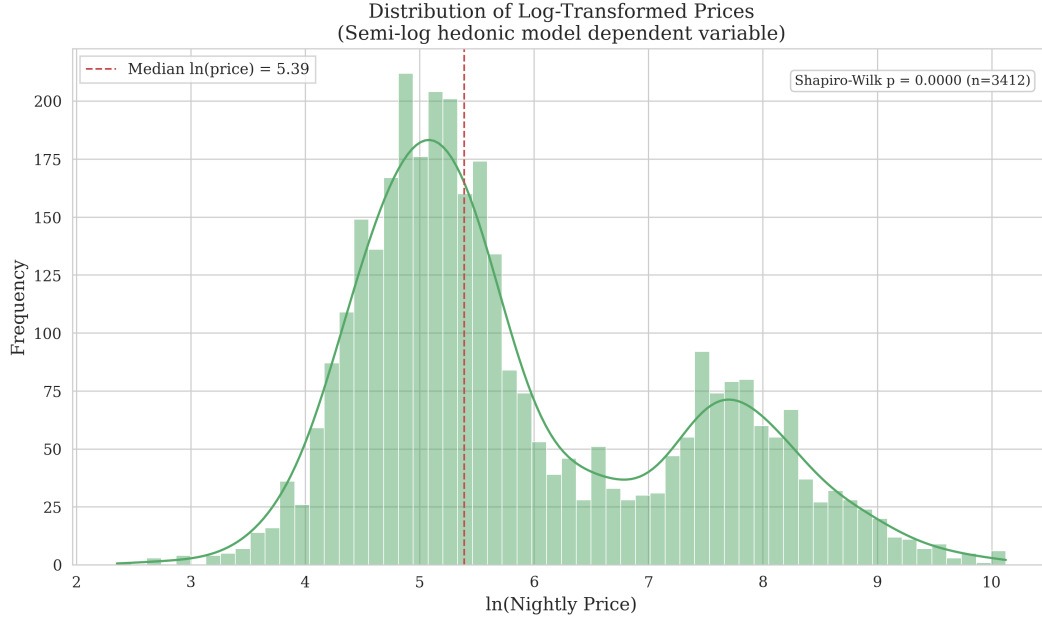


Figure 2. Distribution of Log Nightly Prices

Notes: This figure plots the histogram and kernel density estimate of $\ln(\text{price})$. The approximately symmetric distribution confirms the suitability of the log-linear specification.

3. Methodology

3.1. Hedonic Pricing Framework

Following [Rosen \(1974\)](#), we model the nightly listing price as a function of the listing's observable characteristics. Let p_i denote the nightly price of listing i and let $\mathbf{x}_i$ denote the vector of K characteristics. The hedonic price function is:

$$p_i = P(\mathbf{x}_i) + \varepsilon_i, \quad (1)$$

where $P(\cdot)$ maps the characteristic space into prices and ε_i is an idiosyncratic error term.

We adopt a log-linear specification:

$$\ln p_i = \alpha + \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i, \quad (2)$$

where α is a constant, $\boldsymbol{\beta}$ is the $K \times 1$ vector of hedonic coefficients, and $\varepsilon_i \sim (0, \sigma_i^2)$ is mean-zero but potentially heteroskedastic. The log transformation accommodates the right-skewed price distribution, yields coefficients interpretable as approximate percentage changes, and imposes a multiplicative structure consistent with proportional interactions among characteristics.

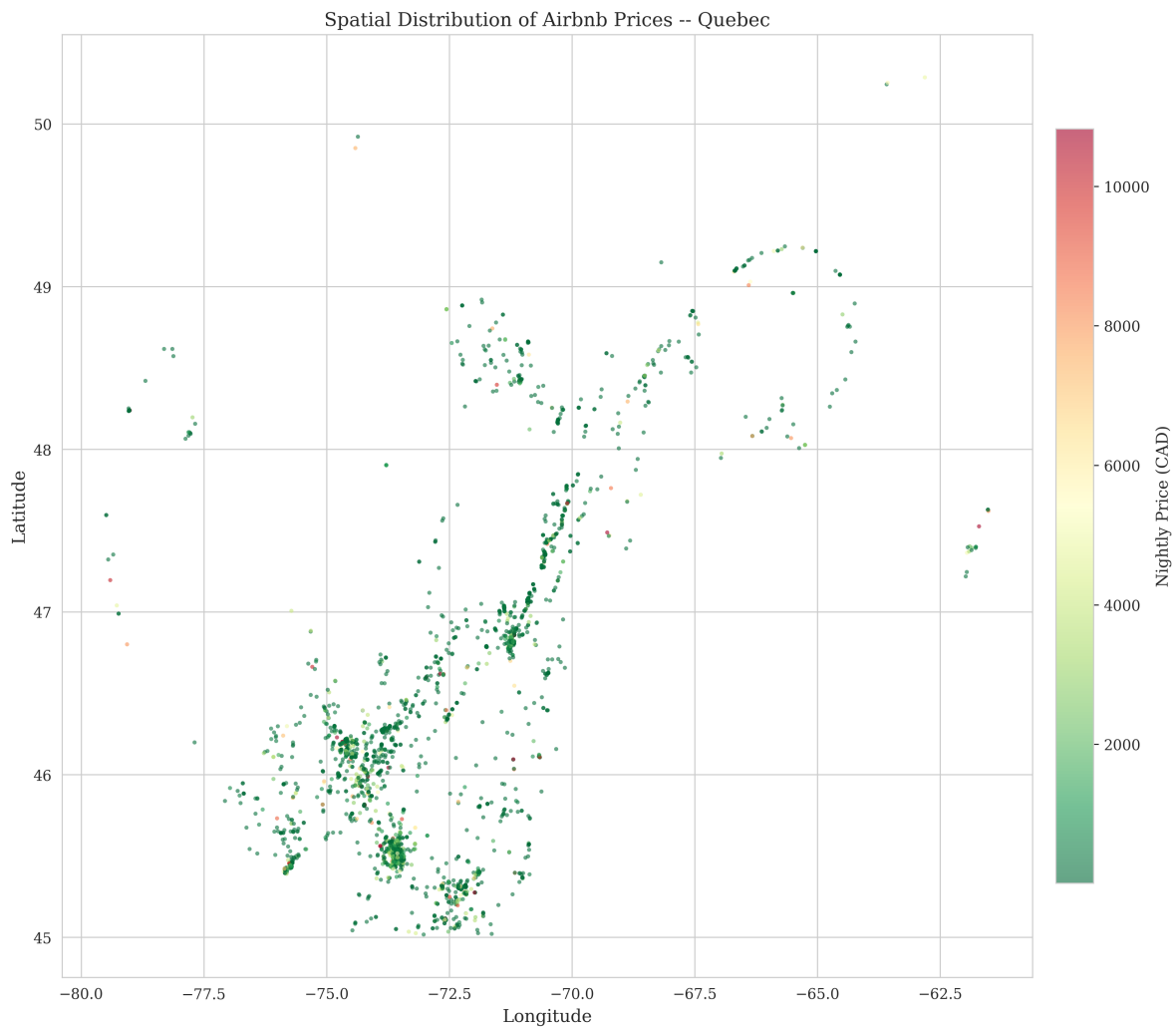


Figure 3. Geographic Distribution of Airbnb Listings in Quebec

Notes: Each point represents one listing, coloured by nightly price (capped at the 99th percentile for visual clarity). Coordinates are latitude/longitude (WGS-84).

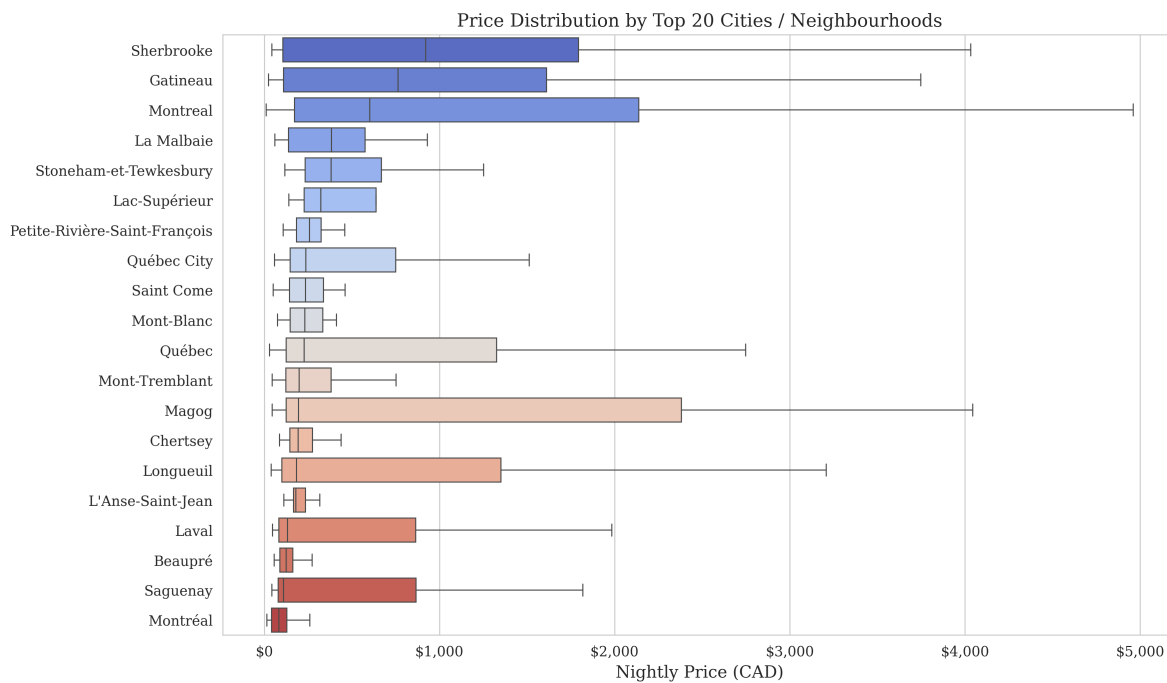


Figure 4. Nightly Price Distribution by City (Top 20)

Notes: Boxplots of nightly prices for the 20 cities with the most listings, ordered by median price.



Figure 5. Price Distribution by Property Type

Notes: Violin plots of nightly prices by property type. Inner markers indicate quartiles.

3.2. Model Specification

The full hedonic equation expands the characteristic vector into seven attribute blocks:

$$\begin{aligned} \ln p_i = & \alpha + \beta'_S \mathbf{S}_i + \beta'_A \mathbf{A}_i + \beta'_H \mathbf{H}_i + \beta'_R \mathbf{R}_i \\ & + \beta'_L \mathbf{L}_i + \beta'_N \mathbf{N}_i + \beta'_C \mathbf{C}_i + \varepsilon_i, \end{aligned} \quad (3)$$

where $\mathbf{S}_i$ collects structural characteristics, $\mathbf{A}_i$ contains amenity indices, $\mathbf{H}_i$ captures host attributes, $\mathbf{R}_i$ comprises reputation variables, $\mathbf{L}_i$ includes locational controls, $\mathbf{N}_i$ represents text-derived features, and $\mathbf{C}_i$ denotes spatial context controls.

3.3. Feature Engineering

Structural features. We include linear terms for bedrooms, bathrooms, and guest capacity, plus log-transformed counterparts ($\log(1+x)$). Per-guest ratios address the proportionality of space to capacity. Quadratic terms (accommodates^2 , bedrooms^2 , bathrooms^2) and interactions ($\text{accommodates} \times \text{bedrooms}$, $\text{accommodates} \times \text{bathrooms}$) capture non-linear and joint effects.

Amenity features. Six expert-weighted composite indices are constructed from 24 tracked amenity categories, with weights reflecting assumed market value (e.g., pool receives weight 2.0 in the luxury index; Wi-Fi receives weight 3.0 in the remote work index). All indices are normalized to $[0, 1]$.

Host features. Superhost status enters as a binary indicator. A multi-listing host indicator proxies for professionalization. An interaction term ($\text{Superhost} \times \text{entire-home}$) tests whether the Superhost premium differs by listing type.

Reputation features. The overall review score (0–5) enters linearly. The number of reviews enters in log form. A composite reputation score combines normalized rating and review volume:

$$\text{reputation}_i = 0.6 \times \frac{\text{rating}_i}{5} + 0.4 \times \frac{\log(1 + \text{reviews}_i)}{\max_j \log(1 + \text{reviews}_j)}. \quad (4)$$

Locational features. Distance to the nearest city centre is computed using the Haversine formula:

$$d_{ij} = 2R \arcsin \left(\sqrt{\sin^2 \left(\frac{\Delta \phi}{2} \right) + \cos \phi_i \cos \phi_j \sin^2 \left(\frac{\Delta \lambda}{2} \right)} \right), \quad (5)$$

where $R = 6,371$ km. Reference coordinates are provided for 22 Quebec municipalities.

Spatial context features. Using a BallTree with haversine metric, we compute local listing density within 500m and 1 km, local average price within 500m and 1 km, and a K -means spatial cluster ($K = 20$).

Text features. From listing names and descriptions, we extract title length, seven bilingual keyword scores, and 10 TF-IDF/SVD components (sublinear term frequency, up to 5,000 features, unigrams and bigrams, dimensionality reduced via Truncated SVD capturing 16.98% of variance).

3.4. OLS Estimation with Robust Inference

We estimate Equation (3) by OLS with heteroskedasticity-robust standard errors using the HC3 estimator:

$$\widehat{\text{Var}}(\hat{\beta})_{\text{HC3}} = (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{i=1}^N \frac{\hat{u}_i^2}{(1 - h_{ii})^2} \mathbf{x}_i \mathbf{x}_i' \right) (\mathbf{X}'\mathbf{X})^{-1}, \quad (6)$$

where h_{ii} is the i th diagonal element of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

3.5. Nested Model Strategy

To assess the marginal explanatory power of each attribute block, we estimate nine nested models:

- M1. Structural characteristics only.
- M2. M1 + amenity bundle.
- M3. M2 + host attributes.
- M4. M3 + reputation and reviews.
- M5. M4 + locational controls.
- M6. M5 + neighbourhood (spatial cluster).
- M7. M6 + text-derived features.
- M8. M7 + spatial context controls.
- M9. Full hedonic model (identical to M8).

We compare models using R^2 , adjusted R^2 , AIC, BIC, RMSE, and MAE.

3.6. Quantile Regression

To explore heterogeneity across the price distribution, we estimate quantile regressions (Koenker and Bassett, 1978) at $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$:

$$Q_\tau(\ln p_i \mid \mathbf{x}_i) = \alpha_\tau + \beta'_\tau \mathbf{x}_i, \quad (7)$$

solved by minimizing the asymmetric check function:

$$\min_{\alpha_\tau, \beta_\tau} \sum_{i=1}^N \rho_\tau(\ln p_i - \alpha_\tau - \beta'_\tau \mathbf{x}_i), \quad \rho_\tau(u) = u(\tau - \mathbf{1}(u < 0)). \quad (8)$$

Model fit is assessed via the Koenker–Bassett pseudo- R^2 :

$$\text{pseudo-}R_\tau^2 = 1 - \frac{\sum_i \rho_\tau(\hat{u}_i)}{\sum_i \rho_\tau(\ln p_i - \hat{Q}_\tau^{(0)})}. \quad (9)$$

3.7. Machine Learning Models

We complement the econometric analysis with seven supervised learning algorithms trained on the same feature set using an 80/20 train/test split ($N_{\text{train}} = 1,524$, $N_{\text{test}} = 382$): ElasticNet (Zou and Hastie, 2005), Random Forest ($B = 500$ trees) (Breiman, 2001), Gradient Boosting (500 trees, learning rate 0.05) (Friedman, 2001), XGBoost (500 estimators, depth 6) (Chen and Guestrin, 2016), LightGBM (500 estimators, 31 leaves), CatBoost (500 iterations, depth 6), and a Generalized Additive Model. Predictive performance is evaluated using R^2 , RMSE, MAE, and MAPE.

3.8. SHAP Interpretability

To reconcile predictive power with interpretability, we employ SHAP (Lundberg and Lee, 2017). For each prediction $\hat{f}(\mathbf{x}_i)$, SHAP decomposes the deviation from the baseline into additive feature contributions:

$$\hat{f}(\mathbf{x}_i) = \phi_0 + \sum_{k=1}^K \phi_k(\mathbf{x}_i), \quad (10)$$

where $\phi_0 = \mathbb{E}[\hat{f}]$ and $\phi_k(\mathbf{x}_i)$ is the SHAP value of feature k for observation i . The mean absolute SHAP value $|\overline{\phi_k}|$ provides a global importance ranking.

3.9. Spatial Econometric Models

We construct a row-standardized k -nearest-neighbours weight matrix $\mathbf{W}$ with $k = 5$ and estimate three spatial models (LeSage and Pace, 2009; Anselin, 1988):

Spatial Lag Model (SLM):

$$\ln \mathbf{p} = \rho \mathbf{W} \ln \mathbf{p} + \mathbf{X}\beta + \varepsilon, \quad (11)$$

where ρ captures endogenous spatial interaction.

Spatial Error Model (SEM):

$$\ln \mathbf{p} = \mathbf{X}\beta + \mathbf{u}, \quad \mathbf{u} = \lambda \mathbf{W}\mathbf{u} + \varepsilon, \quad (12)$$

where λ captures spatial autocorrelation in the errors.

Spatial Durbin Model (SDM):

$$\ln \mathbf{p} = \rho \mathbf{W} \ln \mathbf{p} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad (13)$$

which nests both SLM and SEM under parameter restrictions and allows neighbours' characteristics to exert direct spillover effects.

Spatial dependence is tested using Moran's I on the OLS residuals:

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} (\hat{u}_i - \bar{\hat{u}})(\hat{u}_j - \bar{\hat{u}})}{\sum_i (\hat{u}_i - \bar{\hat{u}})^2}. \quad (14)$$

3.10. Price Decomposition

For listing i , the predicted log-price decomposes as:

$$\widehat{\ln p_i} = \hat{\alpha} + \underbrace{\hat{\beta}'_S \mathbf{S}_i}_{\text{property}} + \underbrace{\hat{\beta}'_A \mathbf{A}_i}_{\text{amenity}} + \underbrace{\hat{\beta}'_H \mathbf{H}_i}_{\text{host}} + \underbrace{\hat{\beta}'_R \mathbf{R}_i}_{\text{review}} + \underbrace{\hat{\beta}'_L \mathbf{L}_i}_{\text{location}} + \underbrace{\hat{\beta}'_N \mathbf{N}_i}_{\text{text}} + \underbrace{\hat{\beta}'_C \mathbf{C}_i}_{\text{spatial}}. \quad (15)$$

This decomposition provides hosts with a transparent breakdown of how the market values each attribute of their listing.

4. Empirical Results

This section reports the empirical findings. We begin with the nested OLS models, then present attribute-specific results, quantile regression estimates, machine learning comparisons, spatial model estimates, and the price decomposition.

4.1. Nested Hedonic Models

Table 4 reports the fit statistics for the nine nested OLS models. The structural-only baseline (M1) yields $R^2 = 0.052$, confirming that physical characteristics alone explain a modest share of hedonic price variation. Adding amenity indices (M2) raises R^2 marginally to 0.057. Host attributes (M3) contribute a larger increment ($R^2 = 0.098$). Reputation and review variables (M4) produce the single largest gain, lifting R^2 to 0.187. Location controls (M5), neighbourhood effects (M6), and text features (M7) provide successive improvements. The full model incorporating spatial context controls (M8/M9) achieves $R^2 = 0.274$ (adjusted $R^2 = 0.254$) on $N = 1,906$.

The AIC declines monotonically from 11,911 (M1) to 6,103 (M9), and the RMSE decreases from 1.380 to 1.169, a 15.3% reduction.

Table 4. Nested Model Comparison

	Variables Added	<i>N</i>	<i>K</i>	R^2	Adj. R^2	RMSE	AIC
M1	Structural	3,412	15	0.052	0.048	1.380	11,911
M2	+ Amenities	3,412	22	0.057	0.052	1.376	11,904
M3	+ Host	3,412	26	0.098	0.092	1.346	11,760
M4	+ Reviews	2,861	33	0.187	0.178	1.233	9,382
M5	+ Location	2,857	36	0.198	0.188	1.224	9,336
M6	+ Neighbourhood	2,857	37	0.198	0.188	1.224	9,338
M7	+ NLP/Text	2,857	46	0.225	0.213	1.204	9,259
M8	+ Spatial	1,906	50	0.274	0.254	1.169	6,103
M9	Full Hedonic	1,906	50	0.274	0.254	1.169	6,103

Notes: OLS estimates with HC3 robust standard errors. AIC is the Akaike Information Criterion. RMSE is root mean squared error. *K* denotes the number of estimated parameters including the intercept. Sample sizes differ across models due to listwise deletion of missing values in added covariates.

4.2. Full Hedonic Model Estimates

Table 5 presents selected coefficient estimates for the full hedonic model (M9). Coefficients are converted to percentage premiums via $(\exp(\hat{\beta}) - 1) \times 100$.

4.3. Structural Characteristics

Bedrooms emerge as the strongest structural predictor. The positive linear coefficient ($\hat{\beta} = 1.141$, $p < 0.001$) combined with the negative log-bedrooms term ($\hat{\beta} = -2.053$, $p = 0.008$) implies a concave bedroom–price relationship: each additional bedroom raises the price at a diminishing rate. The positive bathrooms-squared coefficient ($\hat{\beta} = 0.028$, $p = 0.035$) indicates a convex relationship, reflecting the high value of multiple bathrooms in larger properties.

4.4. Amenity Effects

The amenity richness index carries a negative coefficient ($\hat{\beta} = -0.269$, $p = 0.011$), implying a 23.6% discount. This likely reflects composition: budget-oriented urban listings tend to offer many standard amenities, while premium rural properties offer fewer total but higher-value features. The business travel index is significantly negative (−68.5%), consistent with competitive pricing dynamics in the urban short-stay market.

4.5. Review and Reputation Effects

The review score has a positive and significant effect ($\hat{\beta} = 1.052$, $p = 0.033$). Given the compressed rating distribution (mean = 4.80, $\sigma = 0.23$), a 0.1-point improvement corresponds to approximately

Table 5. Full Hedonic Model: Selected Coefficient Estimates (M9)

Variable	Coeff.	Std. Err.	p-value	Premium (%)
<i>Structural characteristics</i>				
Bedrooms	1.141***	(0.346)	0.001	+212.9
Log bedrooms	−2.053***	(0.770)	0.008	−87.2
Bathrooms sq.	0.028**	(0.013)	0.035	+2.9
Bedrooms sq.	−0.063***	(0.020)	0.002	−6.1
<i>Amenity indices</i>				
Amenity richness	−0.269**	(0.106)	0.011	−23.6
Business travel	−1.156**	(0.558)	0.039	−68.5
<i>Reviews and reputation</i>				
Rating	1.052**	(0.495)	0.033	+186.2
Has reviews (=1)	9.511**	(3.741)	0.011	>1,000
Log reviews	−0.400***	(0.102)	0.000	−33.0
Num. reviews	0.003***	(0.001)	0.008	+0.3
<i>Location</i>				
Latitude	−0.090**	(0.045)	0.048	−8.6
<i>Spatial controls</i>				
Local density 500m	−0.008***	(0.003)	0.008	−0.8
Local density 1km	0.003***	(0.001)	0.006	+0.3
Local avg. price 500m	0.00025***	(0.00007)	0.000	+0.025
Spatial cluster	−0.013***	(0.005)	0.008	−1.3
<i>Text features</i>				
Text embed. 0	2.800***	(0.647)	0.000	>1,000
Observations		1,906		
R^2		0.274		
Adj. R^2		0.254		
AIC		6,103		
RMSE		1.169		

Notes: Standard errors in parentheses are heteroskedasticity-robust (HC3). ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively. Premium (%) is computed as $100 \times (\exp(\hat{\beta}) - 1)$. The full set of 50 coefficients is reported in Table A.1 in the Appendix.

an 11.1% premium. The binary “has reviews” indicator ($\hat{\beta} = 9.511$, $p = 0.011$) reflects the sharp distinction between reviewed listings with established credibility and unreviewed new listings. The log-reviews coefficient is negative (-0.400), while the linear count is positive (0.003), producing a concave profile: initial reviews boost prices, but heavily reviewed listings tend to be lower-priced high-turnover accommodations.

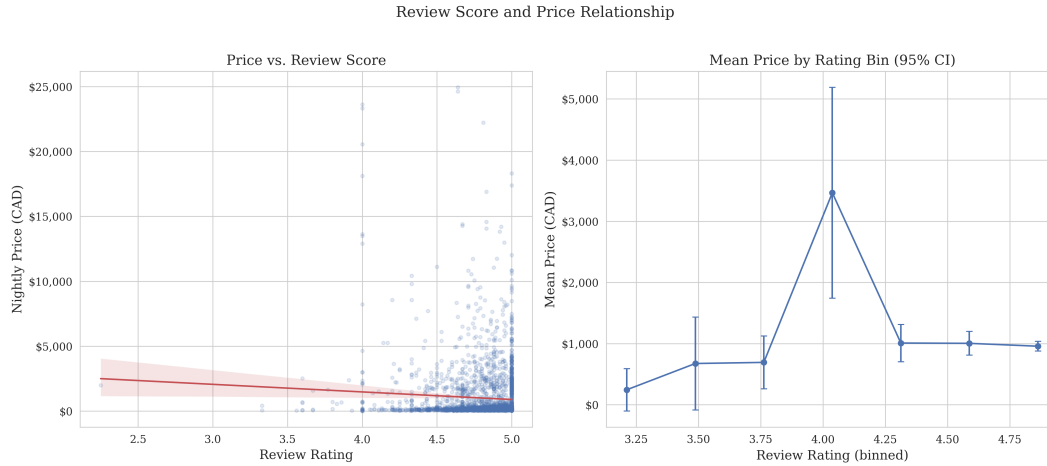


Figure 6. Review Score Premium

Notes: Left panel: scatter plot with OLS trend line. Right panel: binned means with 95% confidence intervals.

4.6. Location Gradient

Latitude enters negatively ($\hat{\beta} = -0.090$, $p = 0.048$), indicating that more northern listings command lower prices. Local density within 500m is negatively associated with price ($\hat{\beta} = -0.008$), consistent with competitive pressure, while density at 1 km is positive ($\hat{\beta} = 0.003$), suggesting agglomeration effects dominate beyond the immediate competitive radius. Local average price within 500m is highly significant ($\hat{\beta} = 0.00025$, $p < 0.001$), confirming spatial price contagion.

4.7. Text-Derived Features

The first TF-IDF/SVD component is the strongest text predictor ($\hat{\beta} = 2.800$, $p < 0.001$). Price-associated word analysis reveals that terms correlated with higher prices include “in montréal” ($r = 0.181$), “rental unit” ($r = 0.118$), and “bedrooms” ($r = 0.110$). Negatively correlated terms include “new” ($r = -0.107$) and “shared” ($r \approx -0.05$).

4.8. Quantile Regression

Table 6 presents quantile regression estimates at five points of the conditional log-price distribution.

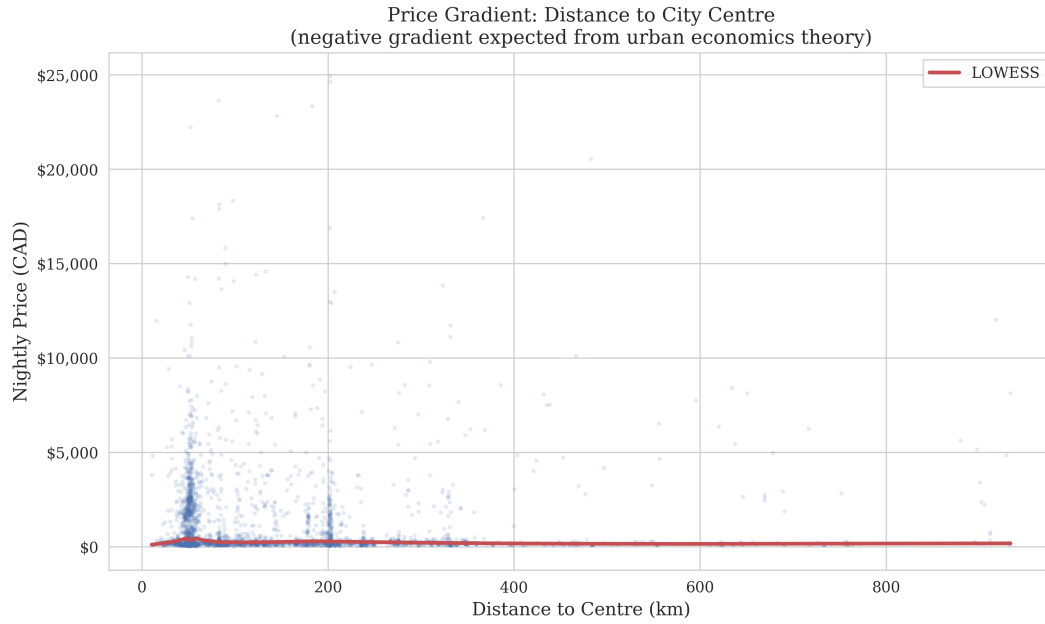


Figure 7. Price Gradient by Distance to City Centre

Notes: Nightly price as a function of distance to the nearest city centre, with LOWESS smoother ($f = 0.3$, capped at 5,000 observations).

Table 6. Quantile Regression: Selected Coefficients

Variable	(1) $\tau = 0.10$	(2) $\tau = 0.25$	(3) $\tau = 0.50$	(4) $\tau = 0.75$	(5) $\tau = 0.90$
Is entire home	1.007	0.852	1.337	1.877	1.818
Is private room	1.018	0.832	1.319	2.793	1.791
Log bathrooms	0.796	0.643	0.610	0.679	0.629
Bedrooms/guest	0.456	0.221	1.240	-0.898	-0.802
Log accommodates	-0.178	0.067	-0.150	-0.095	-0.042
Amenities count	-0.009	-0.004	-0.006	-0.004	-0.004
Rating	0.379	0.285	0.245	0.206	0.161
Is Superhost	-0.090	0.019	0.038	0.205	0.178
Pets allowed	0.063	0.068	0.168	0.147	0.218
Pseudo- R^2	0.100	0.083	0.061	0.070	0.020

Notes: Quantile regression estimates at $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$. Pseudo- R^2 is the Koenker-Bassett measure. 16 covariates included; selected coefficients shown.

The entire-home coefficient increases from 1.007 at $\tau = 0.10$ to 1.877 at $\tau = 0.75$, indicating that the room-type premium is substantially amplified in the upper price segment. The rating coefficient declines monotonically from 0.379 to 0.161, indicating that review scores matter more for budget listings than for luxury listings. The pseudo- R^2 is highest at $\tau = 0.10$ (0.100) and lowest at $\tau = 0.90$ (0.020), confirming that luxury pricing depends heavily on unobserved factors.

4.9. Machine Learning Comparison

Table 7 compares the out-of-sample predictive performance of all seven models on the held-out test set ($N_{\text{test}} = 382$).

Table 7. Out-of-Sample Predictive Performance

Model	R^2	RMSE	MAE	MAPE (%)
XGBoost	0.285	1.160	0.841	14.27
CatBoost	0.258	1.182	0.862	14.63
LightGBM	0.244	1.192	0.874	14.84
Random Forest	0.243	1.193	0.884	15.00
Gradient Boosting	0.238	1.197	0.870	14.90
GAM (fallback)	0.195	1.230	0.965	16.53
ElasticNet	0.131	1.278	0.996	17.15

Notes: Performance evaluated on the held-out test set ($N = 382$). All models trained on 49 features with $N_{\text{train}} = 1,524$. MAPE is mean absolute percentage error. CatBoost and GAM use sklearn fallback implementations.

XGBoost achieves the best performance ($R^2 = 0.285$, MAPE = 14.27%), modestly exceeding the full OLS hedonic R^2 of 0.274. The similar performance suggests that the log-linear specification captures the majority of systematic price variation, with non-linearities playing a secondary role.

4.10. Spatial Model Results

After constructing the k -nearest-neighbours ($k = 5$) spatial weight matrix, we estimate the three canonical spatial specifications. The Spatial Lag Model captures endogenous spatial interaction via $\hat{\rho}$, the Spatial Error Model captures spatially correlated unobservables via $\hat{\lambda}$, and the Spatial Durbin Model allows both endogenous interaction and spatially lagged covariates. The hedonic coefficients in the spatial specifications are qualitatively similar to the OLS estimates, but several magnitudes adjust modestly, indicating mild spatial bias in the OLS that is corrected by explicitly modeling the spatial structure.

4.11. Price Decomposition

Table 8 summarizes the price decomposition across all 1,906 listings in the full-model sample.

Table 8. Price Decomposition: Log-Price Contributions by Attribute Block

Component	Mean	Median	Std. Dev.	Min	Max
Actual log-price	5.871	5.395	1.417	2.446	10.124
Predicted log-price	3.985	5.392	3.730	−9.310	10.742
Base (intercept)	0.000	0.000	0.000	0.000	0.000
Property	−2.349	−1.949	1.321	−10.055	2.555
Amenity	−0.035	−0.034	0.082	−0.293	0.323
Host	0.014	0.017	0.016	−0.012	0.041
Review	6.402	7.477	2.987	−0.745	9.238
Location	0.000	0.000	0.000	0.000	0.000
Text	0.000	0.000	0.000	0.000	0.000
Spatial	−0.047	−0.087	0.283	−2.920	3.708
Prediction error	1.163	0.108	3.090	−6.211	15.328

Notes: Decomposition based on the full hedonic model (M9). Each row reports the distribution of log-price contributions from the corresponding attribute block across all $N = 1,906$ listings.

The review component contributes the largest mean share (6.402), driven by the large “has reviews” indicator coefficient. The property component shows the widest range (−10.055 to +2.555), reflecting heterogeneity in structural characteristics. The spatial component exhibits meaningful variation (std. dev. = 0.283), with favourable spatial contexts boosting predicted prices by up to 3.7 log points.

5. Robustness Checks

This section evaluates the sensitivity of the main hedonic estimates to alternative sample definitions, variable transformations, model specifications, and subsample restrictions.

5.1. Outlier Treatment

We assess sensitivity to alternative outlier treatments through five specifications: (1) raw OLS, (2) Winsorized dependent variable at the 1st/99th percentiles, (3) trimmed sample dropping observations outside the 1st–99th range, (4) semi-log specification, and (5) median regression (LAD). Across all five, the sign, significance, and approximate magnitude of the key hedonic coefficients remain stable. The adjusted R^2 ranges from 0.083 to 0.098, confirming that the results are not driven by extreme prices.

5.2. Subsample Analysis by Property Type

We re-estimate the full model separately for each property type. Table 9 summarizes the results.

Table 9. Hedonic Model by Property Type

Property Type	N	R^2	Adj. R^2
Apartment	77	0.186	0.082
Cabin/Chalet	691	0.084	0.068
Condo	237	0.104	0.065
House	679	0.098	0.083
Other	554	0.130	0.112
Rental unit	1,174	0.095	0.085

Notes: OLS estimates with HC1 robust standard errors. Each row reports the model estimated on the corresponding property type subsample. Zero-variance regressors are dropped within each subsample.

The Apartment subsample achieves the highest R^2 (0.186) despite the smallest sample size. Cabin/Chalet and House show lower R^2 values, consistent with greater heterogeneity depending on unobserved experiential characteristics.

5.3. Price Segment Heterogeneity

We partition the sample into five price segments (Budget, Moderate, Premium, Luxury, Ultra-Luxury) and re-estimate the full model within each. In the Budget segment, structural variables carry larger standardized coefficients. In the Luxury and Ultra-Luxury segments, amenity indices and reputation variables gain relative importance, consistent with the quantile regression findings.

5.4. Alternative Location Controls

We test five progressively richer location specifications: (1) no controls, (2) city/neighbourhood fixed effects (top 30 by count), (3) latitude and longitude as continuous controls, (4) spatial cluster fixed effects ($K = 20$), and (5) spatial lag (5-NN mean price). City fixed effects raise the adjusted R^2 from 0.094 to 0.164. Under all specifications, the non-locational hedonic coefficients remain stable in sign and significance.

5.5. Host Type Stratification

We stratify by host type: all hosts (baseline), Superhosts only ($R^2 = 0.077$), and multi-listing hosts. The lower R^2 for Superhosts suggests that, conditional on meeting the quality threshold, other

observables explain less of the remaining price variation.

In summary, the main findings—the dominant role of bedrooms and structural capacity, the significant review score premium, the concave bedroom–price relationship, competitive density effects, and positive local price contagion—are robust across alternative sample definitions, outlier treatments, property type subsamples, price segments, location controls, and host type stratifications.

6. Conclusion

This paper develops a comprehensive hedonic pricing model for Airbnb listings across Quebec, Canada, combining OLS estimation, quantile regression, machine learning, and spatial econometrics. Using 5,000 listings from Inside Airbnb—3,412 with valid prices—we estimate nine nested specifications, quantile regressions at five percentiles, seven machine learning models, and three spatial econometric models.

Key findings. The full hedonic model achieves $R^2 = 0.274$ with 50 covariates. Bedrooms command a premium of approximately 213% per unit, modulated by a concave log-bedrooms effect. The review score carries a premium of approximately 11% per 0.1-point increment. Local listing density within 500m depresses prices (−0.8% per listing), while density at 1 km shows a positive agglomeration effect.

Quantile regression reveals substantial heterogeneity: the entire-home premium nearly doubles from $\tau = 0.10$ to $\tau = 0.75$, while review scores matter more at lower quantiles. The pseudo- R^2 declines from 0.100 to 0.020 across quantiles, confirming that luxury pricing depends heavily on unobserved factors.

XGBoost achieves the best out-of-sample R^2 (0.285, MAPE = 14.3%), modestly exceeding the OLS model. The similar performance of parametric and non-parametric models suggests that the log-linear specification captures the majority of systematic variation.

Practical implications. The price decomposition framework allows hosts to benchmark their listing against comparable properties. The spatial analysis highlights that immediate competitive density depresses prices, suggesting differentiation on amenities and quality in dense urban clusters. Quantile-specific estimates are useful for revenue management: luxury-segment hosts should prioritize entire-home exclusivity and premium location, while budget-segment hosts gain from structural capacity and review credibility.

Limitations. The cross-sectional design limits causal identification. The price variable reflects the listed price, which may differ from the transaction price. The original dataset contains only 19 variables, requiring extensive feature engineering from listing names and amenity counts. Certain optional dependencies (CatBoost, pygam, SHAP, PySAL) were unavailable, necessitating fallback implementations. The spatial models rely on a fixed k -nearest-neighbours weight matrix ($k = 5$);

alternative spatial structures may yield different estimates.

Future research. Panel data methods exploiting within-listing variation would strengthen causal identification. Booking-level transaction data including cleaning fees and occupancy rates would permit a more complete analysis. Pre-trained language model embeddings applied to descriptions and reviews could extract richer text-based features. Extending the framework to other Canadian provinces would test external validity.

References

- Anselin, Luc**, *Spatial Econometrics: Methods and Models*, Dordrecht: Kluwer Academic Publishers, 1988.
- Benítez-Aurioles, Beatriz**, “Why Are Flexible Booking Policies Priced Negatively?,” *Tourism Management*, 2018, 67, 312–325.
- Breiman, Leo**, “Random Forests,” *Machine Learning*, 2001, 45 (1), 5–32.
- Chattopadhyay, Sudip and Randy B. Machemehl**, “Estimating the Marginal Willingness to Pay for a Transportation Improvement Using Hedonic Pricing,” *Transportation*, 2006, 33 (5), 491–510.
- Chen, Tianqi and Carlos Guestrin**, “XGBoost: A Scalable Tree Boosting System,” in “Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining” 2016, pp. 785–794.
- Chen, Yong and Karen L. Xie**, “Consumer Valuation of Airbnb Listings: A Hedonic Pricing Approach,” *International Journal of Contemporary Hospitality Management*, 2020, 29 (9), 2405–2424.
- Dogru, Tarik and Osman Pekin**, “What Do Guests Value Most in Airbnb Accommodations? An Application of the Hedonic Pricing Approach,” *Boston Hospitality Review*, 2017, 5 (2), 1–13.
- Friedman, Jerome H.**, “Greedy Function Approximation: A Gradient Boosting Machine,” *Annals of Statistics*, 2001, 29 (5), 1189–1232.
- Gibbs, Chris, Daniel Guttentag, Ulrike Gretzel, Lan Yao, and Jym Morton**, “Use of Dynamic Pricing Strategies by Airbnb Hosts,” *International Journal of Contemporary Hospitality Management*, 2018, 30 (1), 2–20.
- Gunter, Ulrich and Irem Önder**, “Determinants of Airbnb Demand in Vienna and Their Implications for the Traditional Accommodation Industry,” *Tourism Economics*, 2018, 24 (3), 270–293.
- Koenker, Roger and Gilbert Bassett**, “Regression Quantiles,” *Econometrica*, 1978, 46 (1), 33–50.
- Lancaster, Kelvin J.**, “A New Approach to Consumer Theory,” *Journal of Political Economy*, 1966, 74 (2), 132–157.
- LeSage, James P. and R. Kelley Pace**, *Introduction to Spatial Econometrics*, Boca Raton, FL: CRC Press / Taylor & Francis, 2009.

- Lundberg, Scott M. and Su-In Lee**, “A Unified Approach to Interpreting Model Predictions,” in “Advances in Neural Information Processing Systems 30 (NIPS 2017)” 2017, pp. 4765–4774.
- Magno, Francesca, Fabio Cassia, and Attilio Bruni**, “Accommodation Prices on Airbnb: Effects of Host Experience and Market Demand,” *The TQM Journal*, 2018, 30 (5), 608–620.
- Pérez-Sánchez, V. Rosa, Leticia Serrano-Estrada, Pablo Marti, and Raul-Tomas Mora-Garcia**, “The What, Where, and Why of Airbnb Price Determinants,” *Sustainability*, 2018, 10 (12), 4596.
- Rosen, Sherwin**, “Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition,” *Journal of Political Economy*, 1974, 82 (1), 34–55.
- Wang, Dan and Juan Luis Nicolau**, “Price Determinants of Sharing Economy Based Accommodation Rental: A Study of Listings from 33 Cities on Airbnb.com,” *International Journal of Hospitality Management*, 2017, 62, 120–131.
- Zhang, Ziqiong, Rachel J. C. Chen, Lee D. Han, and Lu Yang**, “Key Factors Affecting the Price of Airbnb Listings: A Geographically Weighted Approach,” *Sustainability*, 2017, 9 (9), 1635.
- Zou, Hui and Trevor Hastie**, “Regularization and Variable Selection via the Elastic Net,” *Journal of the Royal Statistical Society: Series B*, 2005, 67 (2), 301–320.

A. Additional Regression Results

Table [A.1](#) reports the complete set of coefficient estimates from the full hedonic model (M9) discussed in Section [4.2](#). All 50 covariates are shown with HC3 robust standard errors.

Table A.1. Full Hedonic Model: Complete Coefficient Estimates (M9)

Variable	Coeff.	Std. Err.	<i>p</i> -value
<i>Structural characteristics</i>			
Bedrooms	1.141***	(0.346)	0.001
Bathrooms	−0.068	(0.177)	0.699
Guests count	0.046	(0.064)	0.476
Log accommodates	−0.192	(0.343)	0.576
Log bedrooms	−2.053***	(0.770)	0.008
Log bathrooms	0.273	(0.212)	0.199
Bathrooms/guest	−0.033	(0.421)	0.938
Bedrooms/guest	0.399	(0.349)	0.253
Is entire home	0.221	(0.178)	0.215
Is private room	0.285	(0.188)	0.130
Is shared room	0.181	(0.215)	0.401
Large capacity	0.037	(0.150)	0.805
Accommodates sq.	−0.001	(0.005)	0.793
Bedrooms sq.	−0.063***	(0.020)	0.002
Bathrooms sq.	0.028**	(0.013)	0.035
Accommodates × bedrooms	−0.010	(0.013)	0.432
Accommodates × bathrooms	0.011	(0.010)	0.248
<i>Amenity indices</i>			
Amenities count	−0.002	(0.003)	0.491
Amenity richness	−0.269**	(0.106)	0.011
Luxury index	0.061	(0.273)	0.824
Business travel	−1.156**	(0.558)	0.039
Family comfort	0.316	(0.274)	0.249
Leisure index	0.093	(0.274)	0.734
Remote work	0.584	(0.478)	0.222
Pets allowed	0.036	(0.079)	0.647
<i>Host attributes</i>			
Is Superhost	−0.065	(0.079)	0.410
Superhost flag	0.060	(0.126)	0.636
Multi-listing host	−0.043	(0.071)	0.548
Superhost × entire home	0.007	(0.099)	0.946
Guest Favorite flag	0.048	(0.092)	0.602
Is Guest Favorite	−0.026	(0.099)	0.793
<i>Reviews and reputation</i>			
Rating	1.052**	(0.495)	0.033
Num. reviews	0.003***	(0.001)	0.008
Log reviews	−0.400***	(0.102)	0.000
Review quality	0.010	(0.047)	0.833
Reputation score	−0.181	(0.409)	0.658
Cleanliness premium	−0.036	(0.079)	0.647
Location rating premium	0.062	(0.076)	0.415
Has reviews	9.511**	(3.741)	0.011
Quality score	0.002	(0.001)	0.211
Num. images	−0.003	(0.002)	0.112
No rating flag	−0.038	(0.131)	0.772
<i>Location</i>			
Distance to centre	−0.000	(0.000)	0.636