

## Article

# More than Black Boxes: Machine Learning Models' Capacity to Capture Housing Submarket Patterns

José Rojas-Quiroz \*  and Carlos Marmolejo-Duarte 

Department of Architectural Technology (TA), Centre for Land Valuation Policy (CPSV),  
Barcelona School of Architecture (ETSAB), Universitat Politècnica de Catalunya-BarcelonaTech (UPC),  
08028 Barcelona, Spain; carlos.marmolejo@upc.edu

\* Correspondence: jose.carlos.rojas@upc.edu

## Abstract

Machine learning (ML) models like XGBoost have gained traction in housing valuation as they outperform classical hedonic models in predictive accuracy. However, their black box nature has limited their interpretability, until recent advances like SHAP values enabled deeper insights into variable contributions. This study explores whether XGBoost not only predicts housing prices more accurately but can also help identify latent submarket structures, without imposing predefined spatial or socioeconomic boundaries. Using a raw dataset of 7092 housing listings from Barcelona (reduced to 6527 unique listings and finally to 6112 observations after outlier removal), we employ a sequential methodological framework: first, OLS and Spatial Durbin Models identify statistically significant predictors while accounting for spatial dependence; second, an XGBoost model is trained with these validated variables to extract SHAP values quantifying each attribute's model-attributed contribution to individual predicted prices; third, clustering these SHAP values reveals three distinct submarkets with systematic differences in model-attributed valuation patterns. We validate the resulting segments through five complementary procedures: out-of-sample cluster assignment verifying segment generalizability; nested OLS interaction models that formally test differences in model-attributed valuation patterns across clusters; spatial block bootstrap resampling that assesses cluster stability under geographic subsampling; non-parametric tests against the 2017 cadastral value—an independent administrative proxy for spatial value stratification; and spatial autocorrelation statistics confirming non-random geographic structure. Together, this multi-pronged validation provides convergent evidence that the identified segments capture stable and externally coherent patterns in the model's learned valuation structure. Unlike prior studies focused solely on prediction, our approach bridges ML's technical rigor with econometric interpretability. We argue that this capability may stem from XGBoost's tree-based architecture, which naturally partitions the feature space to accommodate heterogeneous valuation patterns. Future research could extend this framework to more recent tree-based models, further advancing interpretable ML for housing market analysis.



Academic Editor: Kevin H. Zhang

Received: 16 April 2026

Revised: 25 June 2026

Accepted: 2 July 2026

Published: 2 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

**Keywords:** XGBoost; SHAP; cluster; hedonic price model; submarket; valuation; real estate

## 1. Introduction

Since the 1960s, the economic and empirical analysis of urban housing markets has been dominated by equilibrium models that have facilitated the development of theories on residential location [1]. This perspective is complemented by the hedonic housing

price literature, which conceptualizes properties as bundles of physical, locational, and contextual characteristics whose observed prices define an implicit price function over those attributes [2].

Hedonic models are based on two fundamental premises: the existence of implicit markets for each housing attribute, and the presence of a long-term equilibrium price for each of them. However, various studies have shown that the real estate market deviates from theoretical ideals. Persistent heterogeneity in preferences and constraints—together with frictions such as transaction costs, information asymmetries, and behavioral and search limitations—can prevent full arbitrage and allow segmented price structures to persist, giving rise to differentiated housing submarkets [1,3]. In each submarket, the same housing attribute may play a distinct role in price formation. For instance, the presence of a swimming pool may be a strong price differentiator in some segments but a near-standard feature in others, implying systematically different implicit prices across groups.

The identification of housing submarkets has been both a methodological and conceptual challenge [1]. The traditional approach consisted of defining submarkets a priori, through administrative divisions, socioeconomic characteristics, or spatial criteria, and subsequently studying each segment [4,5]. While convenient, such segmentation is often arbitrary and only loosely anchored in the theoretical definition of submarkets as segments with systematically different implicit prices.

In recent years, machine learning (ML) methods have emerged as alternatives that outperform classical hedonic price models, which typically correspond to linear regression models estimated through ordinary least squares (OLS), in terms of predictive capacity and error term reduction [6,7]. Among these, Extreme Gradient Boosting (XGBoost) has demonstrated superior performance in modeling housing prices [8,9]. It is an algorithm that sequentially builds decision trees where each new tree is trained to correct the errors of previous trees, giving rise to what is known as gradient boosting. Since decision trees are hierarchical structures that recursively partition the variable space into regions, they can capture non-linear relationships. This is effective in housing markets, where pricing dynamics arise from complex interactions between multiple property dimensions, including the intrinsic physical characteristics, the built environment of the surrounding area, the socioeconomic context of the neighborhood and its absolute spatial attributes (e.g., proximity to landmarks). However, the “black box” nature of XGBoost has limited its application in contexts where interpretability is crucial. The development of techniques such as SHAP (Shapley Additive Explanations) values has begun to address this limitation, allowing quantification of each variable’s contribution to the final model outcome [10]. Rooted in Shapley values from cooperative game theory, SHAP decomposes each prediction into additive feature attributions; because these attributions can be sensitive to correlated predictors, we detail the computation choices and mitigation steps in Section 3.2.4.

The central hypothesis of this research is that tree-based ML models, such as XGBoost, might capture housing submarket heterogeneity by learning complex, non-linear interactions among attributes, and that SHAP values might help make this latent market structure observable. Building on this central hypothesis, we pose two primary research questions:

- (RQ1) Can clustering on observation-level SHAP values derived from an XGBoost model identify housing submarkets with internally consistent valuation patterns?
- (RQ2) Do the resulting submarkets satisfy independent validation criteria consistent with market-relevant heterogeneity rather than modeling artifacts?

Therefore, the main objective of this paper is to develop a methodology that allows the endogenous identification of housing submarkets by using the patterns captured by an XGBoost model, interpreted through its SHAP values.

For this purpose, we use a dataset comprising 7092 multifamily housing listings from Barcelona, collected in March 2020, reduced to 6527 unique listings after spatial deduplication and further to 6112 observations after outlier removal. These data consist of asking (listing) prices rather than realized transaction prices. In the absence of transaction data, asking prices remain informative about the relative structure of attribute valuations across market segments [11]. The implications of this distinction for interpreting SHAP-derived valuations are discussed in Section 5.

The research method is organized as follows. The first step combines a stepwise OLS model with a Spatial Durbin Model (SDM) to identify a validated, spatially robust set of predictors: the OLS provides an Akaike Information Criterion (AIC)-guided variable selection filter, while the SDM retains only those variables whose effects remain significant after accounting for spatial dependence. Subsequently, we implement an optimized XGBoost model with these variables and extract SHAP values for each observation, quantifying each attribute's contribution to the predicted price. On these SHAP values, we apply *k*-means clustering to identify groups of properties that share homogeneous valuation profiles. The resulting submarkets are then validated through five complementary procedures: (i) out-of-sample cluster assignment verifying segment generalizability; (ii) nested OLS interaction models that formally test whether attribute valuations differ across segments; (iii) spatial block bootstrap resampling that assesses cluster stability under geographic subsampling; (iv) non-parametric tests against the 2017 cadastral value, an independent administrative proxy for spatial value stratification; and (v) spatial autocorrelation analysis confirming non-random geographic structure. This multi-layered validation provides convergent evidence that the identified submarkets capture externally coherent heterogeneity patterns consistent with market segmentation.

The results are consistent with our central hypothesis. They suggest that the XGBoost model can recover meaningful heterogeneity in the Barcelona housing market without prior information about submarket delimitation. Cluster analysis reveals statistically significant differences in how housing attributes are valued across submarkets, particularly in physical and spatial dimensions. These divergences underscore distinct price formation patterns, and are consistent with the model's ability to capture heterogeneous valuation structures.

Unlike previous literature that used ML models primarily for predictive purposes, our results suggest that ML models may achieve more than improve prediction: when paired with interpretability techniques such as SHAP, they can help uncover patterns of heterogeneity in the valuation of attributes across market segments. This finding advances the use of ML in real estate research by indicating that predictive gains may derive not only from modeling non-linearities, but also from adapting to underlying submarket structure. By leveraging SHAP to make this capability analyzable, we establish a bridge between the "black box" of advanced algorithms and the need for theoretical understanding of urban phenomena, opening new avenues for the analysis of complex housing markets.

The remainder of the paper is organized as follows. Section 2 reviews the relevant literature, focusing on OLS models as the primary hedonic pricing method and the subsequent emergence of ML models. In the same section we also discuss the theoretical concept of housing submarkets. Section 3 details the dataset and the methodological steps, which sequentially involve data cleaning, variable selection via OLS and SDM, XGBoost model training, SHAP value extraction, clustering based on SHAP values, and a multi-pronged submarket validation combining OLS interaction models, spatial block bootstrap resampling, external variable tests, and geographic coherence analysis. Section 4 presents the results of the submarket patterns identified, and in Section 5 we discuss them and also point out limitations of our data. Finally, Section 6 concludes by summarizing our key findings and outlining their implications for public policy and future research directions.

## 2. Literature Review

### 2.1. Housing Price Formation and the Hedonic Framework

The analysis of housing prices has traditionally relied on the hedonic pricing framework, which conceptualizes dwellings as bundles of interrelated attributes whose implicit prices are revealed through market transactions [2,12]. Within this framework, the observed price of a dwelling reflects the equilibrium outcome of supply and demand for its constituent characteristics, including physical attributes, locational features, and neighborhood conditions.

OLS has been the dominant estimation method for hedonic models due to its simplicity and interpretability. In this context, regression coefficients are interpreted as marginal implicit prices, providing a clear link between theory and empirical implementation. However, this classical approach relies on strong assumptions that are particularly restrictive in housing markets, such as market homogeneity, linearity, and independence among attributes. These assumptions are increasingly difficult to sustain in urban environments characterized by spatial fragmentation, institutional frictions, and heterogeneous demand.

The relevance of these limitations lies in their implications for both valuation accuracy and economic interpretation. When attribute valuations vary systematically across segments of the market, a single global set of coefficients obscures meaningful heterogeneity and may lead to biased inferences regarding the role of housing characteristics.

### 2.2. Extensions of Hedonic Models to Address Heterogeneity

In response to these limitations, a substantial body of literature has proposed extensions to the classical hedonic model aimed at capturing heterogeneity in housing price formation. Spatial econometric models explicitly incorporate spatial dependence and spatially structured omitted variables through mechanisms such as spatial lags and spatially correlated errors [13,14]. These approaches improve model fit and inference by recognizing that housing prices are embedded in spatial networks rather than isolated observations.

To address spatial heterogeneity in a more direct manner, Geographically Weighted Regression allows regression coefficients to vary across space, thereby relaxing the assumption of parameter stationarity [15]. Similarly, quantile regression models estimate heterogeneous effects across the conditional distribution of prices, implicitly allowing attribute valuations to differ between lower- and higher-priced segments of the market [16].

While these approaches represent important methodological advances, they remain constrained by predefined notions of heterogeneity. Spatial models impose geographic structure *ex ante*, and quantile regression conditions on price alone. As a result, heterogeneity is introduced through modeling choices rather than inferred endogenously from the interaction of housing attributes and market outcomes.

### 2.3. Housing Submarkets as a Structural Response to Heterogeneity

The concept of housing submarkets emerged as a structural response to the empirical failure of unitary market models. Submarkets are commonly defined as groups of dwellings that are close substitutes within the group but imperfect substitutes across groups due to persistent differences in attribute valuations, demand profiles, and locational characteristics [1,17,18].

The persistence of submarkets is attributed to frictions on both the demand and supply sides, including transaction costs, information asymmetries, residential attachment, and the durability of the housing stock. These frictions limit arbitrage across segments, allowing heterogeneous pricing structures to coexist within the same metropolitan area.

Empirical identification of submarkets has followed two main approaches. *A priori* methods rely on administrative or expert-defined boundaries, offering simplicity at the

cost of potential arbitrariness. Data-driven approaches employ statistical techniques such as clustering or principal component analysis to group dwellings based on structural, socioeconomic, and spatial characteristics [19]. Empirical evidence suggests that combining spatial and non-spatial attributes yields more robust submarket delineations than either dimension alone [1,20].

Despite their conceptual appeal, traditional submarket approaches face a trade-off between theoretical relevance and practical applicability. A recurring tension is whether submarkets should be defined as spatially contiguous zones or as non-contiguous groupings: contiguous definitions offer administrative simplicity and tend to perform better in applied valuation contexts [21], whereas non-contiguous clusters may capture patterns that cross geographic boundaries.

#### 2.4. ML Models in Housing Valuation

More recently, ML algorithms have been increasingly applied to housing price prediction as an alternative to parametric hedonic models. Unlike traditional regressions, ML methods allow the functional form to emerge from the data, enabling them to capture complex non-linear relationships and high-order interactions among housing attributes.

Empirical studies consistently show that ML models outperform OLS-based hedonic models in terms of predictive accuracy, particularly in high-dimensional settings [7,8,22,23]. Among these approaches, tree-based ensemble methods, and XGBoost in particular, have demonstrated superior performance across a wide range of housing markets [24].

Recent large-scale evidence reinforces these findings. Hong et al. [25] demonstrate that a Random Forest predictor achieves a mean absolute percentage error below 6% in a dense apartment market, compared to approximately 20% for OLS. Gyger et al. [26] extend this line of work by combining gradient boosting with explicit spatial modeling on over 1.5 million German rental listings, showing that the non-linear component captures complex interactions—particularly between structural attributes and locational characteristics—that linear hedonic models miss.

However, the main limitation of ML models lies in their lack of interpretability. Their predictive success has often been treated as a purely technical improvement, with limited attention to the economic meaning of the patterns they capture. As a result, ML applications in real estate have largely remained confined to forecasting exercises, rather than contributing to a deeper understanding of market structure.

#### 2.5. Interpretability, SHAP Values, and Unresolved Gaps

Hong et al. [25] demonstrate that a Random Forest predictor achieves a mean absolute percentage error below 6% in a dense apartment market, compared to approximately 20% for OLS. Recent advances in model interpretability, particularly SHAP values, have begun to bridge the gap between predictive accuracy and economic interpretation. By decomposing individual predictions into feature-level contributions, SHAP values provide localized, conditional measures of marginal effects that are conceptually aligned with the notion of implicit prices in hedonic theory [10].

Rooted in cooperative game theory, Shapley Additive Explanations (SHAP) values assign each input feature a numerical score that quantifies its contribution to a specific prediction. The underlying logic draws on the concept of Shapley values: the attribution for each feature is computed by averaging its marginal contribution across all possible orderings in which the features could be introduced into the model [10]. This ensures that credit is distributed fairly—no single variable can claim the share of another, and the individual attributions always sum exactly to the gap between the prediction for that observation and the global average prediction. SHAP provides a theoretically unique

solution that satisfies three intuitive requirements: predictions are fully accounted for by the sum of feature contributions; features absent from an observation contribute nothing; and if a feature becomes more influential in a revised model, its attributed score cannot decrease.

Existing studies have used SHAP values to reveal non-linearities and asymmetries in housing attribute valuations, offering insights that are inaccessible to global parametric models [6,27]. Li [28] further demonstrates, through simulation experiments comparing SHAP-explained XGBoost to spatial lag and geographically weighted regression models, that SHAP can recover spatial effects—including both spatial autocorrelation and spatially varying coefficients—at levels of accuracy consistent with dedicated spatial statistical models. This result supports the premise that tree-based models trained on spatially referenced data may internalize spatial structure that local interpretation methods can render explicit. In a similar vein, Gyger et al. [26] employ SHAP to analyze feature importance and non-linear dependencies in a boosting-based hedonic rent model, documenting that usable space alone accounts for approximately 25% of the average prediction impact. Nevertheless, the literature has largely focused on interpreting individual feature effects, rather than on exploiting the collective structure of SHAP patterns across observations to identify latent market segments, which is the gap this study addresses.

As a consequence, a critical gap remains: while ML models demonstrably capture complex heterogeneity in housing markets, their potential to endogenously identify and characterize housing submarkets has not been systematically explored. Existing approaches either impose submarket structures *ex ante* or treat ML models as predictors, leaving unexplored the possibility that submarket segmentation may emerge naturally from the interaction patterns learned by the model.

This study addresses this gap by proposing a methodology that leverages the heterogeneity learned by an XGBoost model, interpreted through its SHAP values, to identify housing submarkets endogenously. In doing so, it reframes ML models not merely as superior predictors, but as tools capable of revealing underlying market structures.

### 3. Materials and Methods

#### 3.1. Dataset

The dataset used in this study comprises information on 7092 multifamily housing units listed for sale in the municipality of Barcelona on the Habitacalia website in March 2020. After removing spatially duplicated observations (listings sharing identical geographic coordinates), 6527 unique listings were retained. Each listing includes geographic coordinates along with attributes related to four dimensions: Physical Characteristics, Socioeconomic Context, Built Environment Features, and Spatial Attributes, as detailed in Table 1. Variables in the Socioeconomic dimension, which uses data from the Spanish Statistical Office (INE) [29,30], and Built Environment one, which uses data from OpenStreetMap [31], refer to a context within a 300-m radius around each listing. For further details on how these variables were constructed, please refer to Supplementary Materials S1. *Variable calculation methods*.

The listings were extracted on 17 March 2020, three days after Spain declared a nationwide state of alarm in response to the COVID-19 pandemic (14 March 2020). Of the 6527 listings, 5943 (91.1%) had their last recorded modification prior to the alarm declaration, indicating that the overwhelming majority of observations reflect pre-lockdown pricing conditions. The temporal implications of this context are discussed in Section 5.

Some variables merit particular attention for their interpretive value. *Primary residences* serves as a proxy for working-class neighborhoods, as many centrally located, higher-income areas have a larger share of dwellings used for tourism accommodation or as second homes for affluent households living elsewhere in Catalonia, reducing the share of primary residences.

**Table 1.** Variables, dimensions, and descriptive statistics ( $n = 6527$ ).

| Dimension     | Variable                                           | Type        | Min.   | Max.      | Mean    | Yes (%) |
|---------------|----------------------------------------------------|-------------|--------|-----------|---------|---------|
| Outcome       | Price (EUR)                                        | Numeric     | 50,260 | 4,200,000 | 475,131 | –       |
| Physical      | Area (m <sup>2</sup> )                             | Numeric     | 30     | 300       | 99.03   | –       |
|               | Age (years)                                        | Numeric     | 0      | 120       | 64.99   | –       |
|               | No. rooms                                          | Numeric     | 1      | 8         | 3.01    | –       |
|               | No. bathrooms                                      | Numeric     | 1      | 6         | 1.65    | –       |
|               | Energy Performance Certificate (EPC; 1 = G, 7 = A) | Categorical | 1      | 7         | 2.33    | –       |
|               | New-build                                          | Boolean     | –      | –         | –       | 5.3     |
|               | Large terrace                                      | Boolean     | –      | –         | –       | 4.0     |
|               | Floor level                                        | Numeric     | 1      | 10        | 2.56    | –       |
|               | Lift                                               | Boolean     | –      | –         | –       | 82.3    |
|               | Lift $\times$ Floor level <sup>a</sup>             | Numeric     | –5     | 10        | 1.83    | –       |
|               | A.C.                                               | Boolean     | –      | –         | –       | 30.8    |
|               | Heating                                            | Boolean     | –      | –         | –       | 33.8    |
|               | Building quality (1 = worst, 9 = best)             | Categorical | 1      | 9         | 6.24    | –       |
| Socioeconomic | Unemployment (%)                                   | Numeric     | 0.03   | 0.28      | 0.11    | –       |
|               | Primary residences (%)                             | Numeric     | 0.52   | 0.97      | 0.81    | –       |
|               | Mean household income (EUR/year)                   | Numeric     | 25,860 | 98,221    | 46,814  | –       |
|               | Household size                                     | Numeric     | 1.90   | 3.06      | 2.36    | –       |
|               | High-educated population (%)                       | Numeric     | 0.12   | 0.70      | 0.51    | –       |
| Built env.    | Public transport stops                             | Numeric     | 0      | 37        | 9.73    | –       |
| Spatial       | Dist. Plaça Catalunya (m) <sup>b</sup>             | Numeric     | 79     | 8987      | 2858    | –       |
|               | Dist. Plaça Francesc Macià (m) <sup>b</sup>        | Numeric     | 13     | 8953      | 3022    | –       |
|               | Dist. Seafront (m) <sup>b</sup>                    | Numeric     | 9      | 8358      | 3380    | –       |
|               | Dist. Diagonal Av. (m) <sup>b</sup>                | Numeric     | 0      | 6796      | 1505    | –       |

<sup>a</sup> For this calculation, the value 0 of Lift was converted to –1. <sup>b</sup> Euclidean distance.

We also include straight-line distances to key landmarks: Plaça Catalunya, Barcelona’s historical central reference; Plaça Francesc Macià, a secondary hub with distinct upper-middle-class connotations; Diagonal Avenue, the most recognizable thoroughfare in the city, characterized by commercial and office uses; and the Seafront, reflecting Barcelona’s identity as a coastal city with a prominent beachfront.

### 3.2. Method

The methodology deliberately combines methods of different natures, each serving a specific purpose. Traditional econometric models, such as OLS and the Spatial Durbin Model, identify a validated and spatially robust set of predictors. An XGBoost model then captures the non-linear valuation patterns embedded in those variables, while SHAP values decompose its predictions at the observation level, making latent heterogeneity explicit and quantifiable. A  $k$ -means algorithm groups properties by their SHAP-based valuation profiles, with the aim of surfacing any latent segmentation the model may have internalized. Finally, the resulting segments are validated through a set of complementary procedures (out-of-sample cluster assignment, OLS interaction models, spatial block bootstrap resampling, non-parametric tests, and spatial autocorrelation statistics) that jointly assess whether the identified submarkets exhibit externally coherent patterns consistent with market structure rather than modeling artifacts. The methodological process consists of the following stages.

#### 3.2.1. Variable Preparation and Selection Procedure

Prior to model training, a series of preparatory steps were carried out to reduce multicollinearity, handle sparse variables, and identify a stable and informative feature set.

1. Centrality index via PCA. Distances to Plaça Catalunya, Plaça Francesc Macià, and Avinguda Diagonal were combined into a single centrality index using Principal Component Analysis (PCA), retaining the first principal component (74.20% of variance explained). This reduces multicollinearity among these correlated distance measures, which would otherwise risk distorting SHAP attributions in the subsequent XGBoost model [32].
2. Material conditions index via PCA. Four neighborhood-level variables (*Building quality*, *Population with higher education*, *Unemployment*, and *Mean household income*) were likewise reduced to a single composite index via PCA (81.92% of variance explained). As with the centrality index, the motivation is to reduce multicollinearity that could bias SHAP values.
3. Variables with very low prevalence (*New-build* and *Large terrace*) were excluded. Sparse binary features can produce underrepresented regions in the decision tree feature space, leading to unreliable SHAP values when the model marginalizes over these dimensions.
4. Removal of a highly collinear variable. The *Heating* variable was excluded due to its high correlation with *A.C.* In Barcelona's climate, *A.C.* was judged the more relevant variable to retain. Combining both in a PCA component was considered interpretively forced given their limited conceptual overlap.
5. Outliers in the log-transformed price were identified and removed using the interquartile range (IQR) method with  $k = 1.5$ , reducing the sample from 6527 to 6451 observations. The log transformation is applied throughout all models for two reasons: it renders the distribution of residuals closer to normality, which is a key assumption of OLS-based models; and it yields a semi-logarithmic functional form in which each regression coefficient—and, by extension, each SHAP value—can be interpreted directly as a proportional effect on price.
6. Multivariate outlier detection was then applied using the Mahalanobis distance. Observations whose squared Mahalanobis distance exceeded the 99th percentile of the  $\chi^2$  distribution (with degrees of freedom equal to the number of variables) were removed, yielding a final clean dataset of 6112 observations. Unlike the preceding IQR step, which flags properties with extreme values on price alone, the Mahalanobis distance detects properties that are unusual in the joint combination of their characteristics—for instance, a very large flat at an atypically low price in a high-income neighborhood—that would not be identified by examining any single variable in isolation. An examination of the price distribution of the 339 removed observations shows that 55.8% fall in the top tercile of unit price. This indicates that the Mahalanobis step mainly removed atypical luxury-tier attribute combinations, so that the retained High-End cluster is composed of representative premium properties rather than statistical outliers that could distort cluster centroids.  
Table 2 summarizes the sequential data cleaning steps and their cumulative impact on sample size.
7. A stepwise OLS model was estimated using AIC-based forward-backward selection, using the variables retained after the filtering steps described above but forcing the inclusion of *Area* and its squared term to capture diminishing marginal returns, explicitly accepting the multicollinearity this introduces. Variables retained by the procedure were then screened for multicollinearity using the Variance Inflation Factor (VIF); only those with  $VIF < 3$  were kept. The AIC penalizes model complexity proportionally to the number of parameters added, so each variable is retained only if the improvement in fit exceeds the cost of an additional parameter. This makes the stepwise OLS a statistically grounded filter for discarding uninformative predictors

before the more computationally demanding spatial and ML stages; it is not used for inference or interpretation.

8. The variables retained by the stepwise procedure were further refined by estimating a SDM. This second filter ensured that only variables whose effects were statistically robust after accounting for spatial dependence entered the final feature set, directly addressing the spatial autocorrelation inherent in urban housing data. Only those variables with statistically significant direct or indirect spatial effects were retained. We used a  $k$ -nearest neighbors spatial weights matrix, with  $k = 15$  ensuring a fully connected spatial network while still preserving the local nature of spatial interactions.
9. The variables identified in Step 8 formed the basis for the XGBoost model, with the exception of the squared Area term, which was excluded since tree-based models capture non-linear and diminishing-return relationships directly from the raw variable without requiring explicit polynomial transformations.

**Table 2.** Data cleaning and outlier removal workflow.

| Step                 | n Before | n Removed | n After | Removal % |
|----------------------|----------|-----------|---------|-----------|
| Initial dataset      | –        | –         | 7092    | –         |
| Duplicate removal    | 7092     | 565       | 6527    | 8.0%      |
| IQR outliers (price) | 6527     | 76        | 6451    | 1.2%      |
| Mahalanobis outliers | 6451     | 339       | 6112    | 5.3%      |
| Final dataset        | –        | 980       | 6112    | 13.8%     |

### 3.2.2. Train–Test Split with Spatial Awareness

The cleaned dataset is first randomly divided into a training set (70 percent) and a test set (30 percent) using a fixed random seed for reproducibility. This random holdout test set provides the primary estimate of predictive performance on unseen observations.

Within the training set, five spatial blocks were created via  $k$ -means clustering on geographic coordinates. These blocks served two purposes: as folds for a complementary spatial cross-validation evaluation, which provides a stricter test of predictive generalization by holding out entire geographic zones, and as the geographic units for spatial block bootstrap resampling in the cluster stability analysis (Section 3.2.6). In the spatial cross-validation procedure, entire geographic blocks were held out in each fold, simulating the more demanding scenario of predicting prices in areas not represented in the corresponding training fold.

### 3.2.3. XGBoost Training and Hyperparameter Optimization

Hyperparameters were tuned via Optuna [33], a Bayesian search framework, exploring learning rate, maximum tree depth, minimum child weight, subsample and column-sample ratios, regularization parameters, and number of boosting rounds. Unlike exhaustive grid search, which tests all pre-specified combinations, or random search, which samples blindly, Bayesian optimization builds a probabilistic model of the relationship between hyperparameter configurations and validation performance, using it to prioritize promising regions of the search space and converge on good solutions substantially more efficiently for the same number of trials. The tuning objective was standard 5-fold cross-validation within the training set. After selecting the best performing configuration, model performance was additionally evaluated using spatial cross-validation, where entire spatial blocks were held out in each fold. This complementary evaluation disentangles predictive accuracy under random sampling from the more demanding task of spatial generalization.

### 3.2.4. SHAP Value Extraction and Interpretation

Building on the framework introduced in Section 2, SHAP values were computed here at the individual property level: for each observation  $i$  and feature  $j$ ,  $\phi_{ij}$  quantifies the contribution of feature  $j$  to the predicted price relative to the global average prediction. Their individual attributions sum exactly to the gap between that property's predicted price and the sample mean:

$$\hat{y}_i = \bar{y} + \sum_{j=1}^N \phi_{ij}, \quad (1)$$

where  $\hat{y}_i$  is the predicted log-price and  $\bar{y}$  is the average predicted log-price across all observations. The sum runs over all  $N$  resulted features (which results from Section 3.2.1).

Two properties may have similar characteristics yet very different SHAP profiles; for example, *Area* might contribute +0.30 log units to the price of a compact, entry-level property but only +0.10 for a high-end property, even if both have the same surface area, reflecting heterogeneous model-attributed valuation patterns across segments. This individual sensitivity signature underpins the clustering step: properties with similar SHAP profiles respond similarly to market features, even if their absolute prices differ.

SHAP values were computed using TreeExplainer [34], an exact algorithm that exploits the recursive tree structure of XGBoost to compute Shapley values in polynomial time rather than the exponential time required by model-agnostic approaches. We adopted the path-dependent feature perturbation variant [35], which restricts the marginalization of each feature to the observations that followed the same decision path through the tree during training. This means that SHAP contributions are anchored to feature combinations that actually occurred in the data, rather than to synthetic combinations generated by the marginal distribution—an important distinction when features are correlated, because the interventional variant may query the model at points in the feature space that were sparsely or never populated during training [34].

The tradeoff is that path-dependent attributions may absorb some of the effect of correlated features, since the tree structure reflects the joint distribution of the training data. This risk was partially mitigated by the variable preparation steps described in Section 3.2.1: correlated distance variables are collapsed into a single centrality index via PCA, and correlated neighborhood-level socioeconomic variables are similarly reduced to a composite index, limiting the dependence pathways that could otherwise distort attributions [32]. Variables with very low prevalence (*New-build* and *Large terrace*) were additionally excluded, reducing the risk of underrepresented regions in the decision tree feature space.

SHAP values here should accordingly be interpreted as measures of the model's learned valuation structure, conditioned on the feature combinations present in the training data. For the purpose of clustering, what matters is the relative structure of SHAP profiles across observations rather than the causal purity of each individual attribution: profiles that are similar in structure will remain similar after path-dependent computation, preserving the geometric relationships that  $k$ -means exploits. The subsequent validation framework provides a further safeguard: if residual SHAP unreliability dominated the analysis, we would expect clusters unstable under bootstrap resampling, poor generalization to test data, no coherence with external variables, and purely spatial or random geographic patterns.

### 3.2.5. Clustering Based on SHAP Values

The clustering procedure was performed exclusively on the SHAP value matrix obtained from XGBoost. Each observation was embedded in a multidimensional space where each dimension corresponds to the SHAP value of a given explanatory variable, capturing its attributed contribution to predicted housing price for that specific observation, con-

ditional on the feature combinations present in the training data. Clustering therefore grouped listings according to similarities in valuation patterns, rather than similarities in raw attributes, spatial proximity, or prices themselves.

SHAP values were not standardized prior to clustering. Because the dependent variable is log-price, each SHAP value represents an approximate proportional contribution to predicted housing price. Clustering directly on unstandardized values therefore grouped properties whose price formation is similar in both structure and magnitude—which is consistent with the economic definition of a submarket as a segment with systematically different price-formation patterns. Standardizing would eliminate inter-feature magnitude differences that are substantively meaningful: for instance, location may contribute +0.80 log units to the predicted log-price of a luxury property yet only +0.10 for an entry-level unit—a distinction that is central to submarket identification. Removing such differences would conflate segments whose valuation patterns differ not in kind but in degree, obscuring the very heterogeneity the clustering is intended to surface. We acknowledge that this choice introduces a risk: if *Area* dominates the SHAP space, clusters may primarily reflect size-based groupings rather than genuine submarket structures. This risk is precisely why the subsequent validation procedures are designed to assess whether the resulting segments capture market heterogeneity beyond what a single dominant feature can explain (see Section 5).

K-means was selected as the clustering algorithm. It minimizes within-cluster variance using Euclidean distance, which is coherent with the objective of grouping observations whose unstandardized SHAP profiles are similar in both structure and magnitude. In practice, the algorithm partitions observations into  $k$  clusters by iteratively assigning each observation to the nearest cluster centroid and updating centroids until convergence. Because Euclidean distance treats all directions in the feature space symmetrically,  $k$ -means implicitly favors clusters with approximately equal variance in every direction—i.e., spherical (isotropic) clusters—and partitions the space into convex Voronoi cells bounded by hyperplanes equidistant between pairs of centroids. The optimal number of clusters ( $k$ ) was determined using the elbow method and the silhouette method, both of which indicated  $k = 3$  as the preferred solution (see Section 4).

With only  $k = 3$  clusters in an 11-dimensional space, the spherical-cluster assumption is substantially less restrictive than it would be with many clusters: 3 convex Voronoi regions can accommodate a wide variety of data configurations, and the assumption becomes practically limiting only when numerous clusters must tile a complex feature space. This constraint is transparent and can be evaluated through sensitivity checks. The main alternative approaches present interpretive or practical difficulties in this context: Gaussian Mixture Models (GMMs) assume multivariate Gaussian distributions whose validity in a space shaped by tree partitioning structure rather than a data-generating process lacks theoretical or empirical support, and density-based methods such as DBSCAN require two parameters—a neighborhood radius and a minimum number of points within that radius—that jointly define what counts as a “dense” region. In geographic or standardized feature spaces these thresholds have intuitive interpretations (e.g., a radius in meters or in standard deviations), but in SHAP space each dimension represents a model-attributed contribution to log-price, and there is no established theoretical or empirical criterion for what constitutes a meaningful density threshold. Moreover, DBSCAN classifies low-density observations as noise, which lacks a clear economic interpretation: in a housing market context, every property belongs to some market segment. An empirical comparison with four GMM variants confirmed that the spherical GMM—geometrically analogous to  $k$ -means—yields near-perfect agreement with the  $k$ -means solution (ARI = 0.941); although more flexible GMM variants achieve bet-

ter BIC, they yield substantially lower silhouette scores, indicating less compact and well-separated clusters. Full GMM results are reported in Supplementary Materials S2. *Clustering method sensitivity.* GMM fitted with 10 random initializations and convergence tolerance  $10^{-4}$ . Silhouette scores:  $k$ -means = 0.324, GMM-spherical = 0.324, GMM-tied = 0.198, GMM-full = 0.125, GMM-diagonal = 0.134. Cluster characterization was based on average price per square meter associated with each SHAP-based group, not on price information used during clustering.

### 3.2.6. Cluster Stability Analysis (Spatial Block Bootstrap)

To assess whether the  $k$ -cluster solution is robust or merely an artifact of the particular training sample, we performed spatial block bootstrap resampling. Unlike standard bootstrap that resamples individual observations, spatial block bootstrap resamples entire geographic blocks with replacement, preserving within-block spatial autocorrelation structure—a more appropriate strategy when spatial dependence is present [36,37].

In each of  $B = 1000$  bootstrap iterations [38,39], we drew a bootstrap sample of five blocks with replacement from the original five spatial blocks, so that, on average, one block was absent and served as the out-of-bag (OOB) set, and collected all in-bag observations.

A new  $k$ -means model was then fitted on the pre-computed SHAP values of those in-bag observations. Cluster assignments were predicted for the OOB observations and compared to their original cluster labels using the Adjusted Rand Index (ARI). This OOB evaluation strategy avoids inflating stability estimates by assessing cluster reproducibility exclusively on observations that played no role in fitting the bootstrap  $k$ -means model. ARI ranges from 0 (random labeling) to 1 (perfect agreement), with values above 0.80 generally considered evidence of stable clustering [40]. In practical terms, each iteration answers the question: when the model is retrained on a geographically different subsample of the city, do the held-out properties receive the same cluster assignments as in the original solution? The ARI quantifies this by considering all pairs of held-out observations and measuring the proportion of pairs that are consistently assigned to the same cluster—or consistently kept in different clusters—across the two solutions, corrected for the level of agreement expected purely by chance.

### 3.2.7. Test Data Cluster Assignment

This step guards against the artifact of data leakage: if test-set observations had contributed to centroid estimation, apparent generalizability would reflect contamination rather than genuine out-of-sample market structure. The clustering model was therefore trained exclusively on the training set. To verify that segments generalize to new data, each test set observation was assigned to the cluster with the nearest centroid in SHAP space (Euclidean distance). This ensured that test set properties were classified based on their SHAP profiles without any influence from test set information during the clustering process. Comparable cluster proportions between training and test sets served as the criterion for segment generalizability.

### 3.2.8. OLS Interaction Model

A fundamental criticism of SHAP-based segmentation may be potential circularity: if segments are identified using a price prediction model and then shown to predict prices, the analysis may simply be confirming model self-consistency rather than discovering genuine market structures.

The OLS specifications estimated here are themselves inherently subject to this circularity: cluster assignments derive from SHAP values, which in turn derive from model predictions, so any systematic pattern the model has captured—whether genuine market heterogeneity or a modeling artifact—will produce clusters that explain additional vari-

ance in prices. Three nested OLS specifications were therefore designed to assess whether observed patterns are statistically systematic and replicable on the held-out sample, rather than to resolve the circularity. This out-of-sample parameter-stability approach aligns with established validation practices in housing submarket research [3]; its results are intended to be read in conjunction with the two subsequent steps—external variable validation and geographic coherence—which rely on information entirely independent of the model and thereby break the circular dependence on model-derived quantities.

Model 1 (main effects) included the 11 predictor variables and two cluster dummy variables, using Entry-Level (the lowest-price cluster) as the reference category. It assessed whether cluster membership retains explanatory power for prices after conditioning on all observable property characteristics. Two complementary tests were applied to the cluster dummies: individual significance tests on each coefficient, and a joint Wald test on the pair. The Wald test evaluates whether the two cluster dummies are jointly significant—that is, whether taken together they account for statistically systematic variation in prices beyond what the eleven predictors already explain. If SHAP-derived segments merely recapitulated what the predictors already explain, neither the individual coefficients nor the joint test would reach significance.

Model 2 (interaction effects) augmented Model 1 by adding full interactions between all 11 explanatory variables and each of the two non-reference cluster dummies (Mid-Range and High-End), testing for differential attribute valuation across segments:

$$\ln(P_i) = \beta_0 + \sum_k \beta_k C_{ik} + \sum_j \gamma_j X_{ij} + \sum_{j,k} \delta_{jk} (X_{ij} \times C_{ik}) + \epsilon_i, \quad (2)$$

where  $P_i$  is the observed price of dwelling  $i$ ;  $C_{ik}$  denotes a dummy variable indicating whether property  $i$  belongs to cluster  $k$ ;  $X_{ij}$  are the selected explanatory variables;  $X_{ij} \times C_{ik}$  represents interaction terms between housing attributes and cluster membership;  $\delta_{jk}$  captures the differential effect of characteristic  $j$  in cluster  $k$  relative to the reference group; and  $\epsilon_i$  is a scalar error term assumed independent and identically distributed with  $\mathbb{E}(\epsilon_i) = 0$  and  $\text{Var}(\epsilon_i) = \sigma^2$ . Model 1 corresponds to the restricted case where all  $\delta_{jk} = 0$ . Interactions and intercept shifts associated with the Entry-Level cluster are absorbed into the reference category to avoid multicollinearity.

The comparison between Models 1 and 2 proceeded at two levels. First, a Wald test assessed whether the full set of interaction terms in Model 2 (all predictor variables  $\times$  non-reference clusters) was jointly significant. In plain terms, this asks: taken together, is there statistically significant evidence that housing attributes are priced differently across segments, beyond what sampling variability alone could produce? A significant result means the interaction terms collectively capture real structure rather than noise, providing an omnibus test of differential attribute valuation across submarkets. Second, individual interaction coefficients were examined to identify which specific attribute–cluster combinations drove the heterogeneity.

Model 3 re-estimated the same interaction specification as Model 2 but with Mid-Range as the reference category, replacing Entry-Level. This re-parameterization did not change the underlying model—fit statistics and the joint significance of the interaction block are identical—but it made the Entry-Level–Mid-Range and Mid-Range–High-End contrasts directly readable from the coefficient table, complementing the Entry-Level–Mid-Range and Entry-Level–High-End contrasts reported in Model 2.

### 3.2.9. External Validation Design

This step targets the artifact of circular validation: if clusters are meaningful only within the model’s own prediction space, coherence tests that rely on model-derived

quantities cannot distinguish genuine market heterogeneity from systematic patterns that the algorithm has learned to reproduce. Validation against a variable computed entirely outside the model breaks this circularity. External validation therefore drew on a source entirely independent of the model: the 2017 cadastral value, an administratively defined, market-informed external benchmark aggregated at the cadastral valuation zone level. Cadastral values are administratively determined assessments of land and built property computed by the Dirección General de Catastro (Spanish Ministry of Finance) [41], intended to approximate market values through standardized criteria. If clusters capture patterns consistent with market heterogeneity, they should differ systematically on this independent indicator.

As a preliminary step, Shapiro–Wilk tests were applied to each cluster’s cadastral value distribution to check for normality. If normality is rejected in any group, the parametric one-way ANOVA—which requires approximately normal residuals—is no longer appropriate, and a non-parametric alternative must be used. We then examined whether clusters identified through SHAP profiles aligned with the 2017 cadastral value using the Kruskal–Wallis test. Rather than comparing group means (as ANOVA does), Kruskal–Wallis ranks all observations from lowest to highest regardless of cluster membership, and then tests whether those ranks are systematically concentrated in certain groups—that is, whether some clusters tend to occupy the top of the ranking while others occupy the bottom, more than chance would predict. Pairwise Mann–Whitney *U* comparisons with Bonferroni correction were then applied to identify which specific cluster pairs drove any overall difference.

### 3.2.10. Geographic Coherence Assessment Design

Geographic coherence validation targeted two complementary artifacts. The first is spatial randomness: clusters distributed uniformly across the urban space would lack geographic grounding and be inconsistent with market-based segmentation. The second is pure spatial determination: clusters concentrated entirely within discrete geographic enclaves would indicate that the algorithm has merely recovered location effects, adding nothing beyond a spatial variable. Two tests jointly assessed whether the identified segments fell in the theoretically expected intermediate position between these extremes.

First, a chi-squared test of independence examined the cross-tabulation of clusters by Barcelona neighborhoods. The test evaluated whether cluster membership was randomly distributed across neighborhoods or showed systematic spatial patterns. The intuition is that if cluster membership were spatially random, each neighborhood would contain roughly the same proportions of High-End, Mid-Range, and Entry-Level properties as observed city-wide; a significant  $\chi^2$  statistic indicates that the deviations from this uniform expectation are too systematic to attribute to chance.

Second, Moran’s *I* statistic quantified spatial autocorrelation using a *k*-nearest-neighbors spatial weights matrix. Cluster membership was encoded as an ordinal variable (Entry-Level = 0 < Mid-Range = 1 < High-End = 2), reflecting the price-tier ranking by which the segments are defined, so that Moran’s *I* operates on a substantively meaningful numeric ordering rather than an arbitrary categorical code. Under genuine market segmentation, neighboring properties would be expected to share cluster membership more often than chance. Intuitively, Moran’s *I* asks whether a given property’s cluster assignment tends to resemble those of its nearest geographic neighbors more than would be expected if assignments were randomly reshuffled across the map.

A statistically significant positive Moran’s *I* confirms non-random geographic structure; its magnitude distinguishes the type of spatial patterning. Values approaching 1 would indicate that segmentation is driven primarily by location, with the algorithm adding little beyond a spatial variable; moderate positive values would indicate non-

random geographic structure without pure spatial determination; values close to zero would indicate that segment membership is spatially random, which would be inconsistent with market-based segmentation.

## 4. Results

The analysis is organized sequentially following the methodological steps described in Section 3. All code documenting the process by which the results were obtained is provided in Supplementary Materials S3. *Code notebooks and variable documentation*, for readers wishing to examine the implementation in depth.

### 4.1. Variable Preparation and Selection

The stepwise AIC-based OLS followed by SDM spatial filtering retained eleven predictor variables: *Area*, *No. rooms*, *No. bathrooms*, *A.C.*, *Age*, *EPC*, *Lift × Floor level*, *Unemployment*, *Building quality*, *Distance to the Seafront*, and the *Primary residences*. The composite *Material conditions index*, intended to aggregate *Building quality*, *Population with higher education*, *Unemployment*, and *Mean household income* into a single PCA component, was excluded at the VIF-checking stage that followed stepwise selection: it exceeded the threshold of 3 ( $VIF = 3.29$ ) adopted to guard against multicollinearity. It was therefore replaced by its constituent variables, *Building quality* and *Unemployment*, both of which had acceptable VIF values (2.40 and 2.52, respectively) and proved significant—*Unemployment* directly in the SDM, and *Building quality* through its spatial lag—and were retained directly. Of the remaining two constituents, *Population with higher education* was excluded because its inclusion raised VIF above the threshold, and *Mean household income* was withheld to avoid contaminating the external validation stage, given its expected strong correlation with the 2017 cadastral value used as an independent benchmark (Section 3.2.9).

The *Centrality index* entered the SDM but was non-significant both directly ( $z = -0.118$ ,  $p = 0.906$ ) and through its spatial lag ( $z = -0.207$ ,  $p = 0.836$ ) and was therefore excluded. Its constituent distance variables were subsequently tested individually, but none reached significance; only *Distance to the Seafront*, which had not been included in the PCA, was retained. The spatial lag coefficient of the SDM was highly significant ( $\rho = 0.4401$ ,  $p < 0.001$ ), confirming the presence of spatial dependence and justifying the spatial econometric filter over non-spatial OLS.

Finally, the  $Area^2$  term was retained by the stepwise OLS to capture diminishing marginal returns but was excluded from the XGBoost specification, as gradient-boosted trees capture such nonlinearities endogenously from the untransformed variable.

### 4.2. XGBoost Model Performance

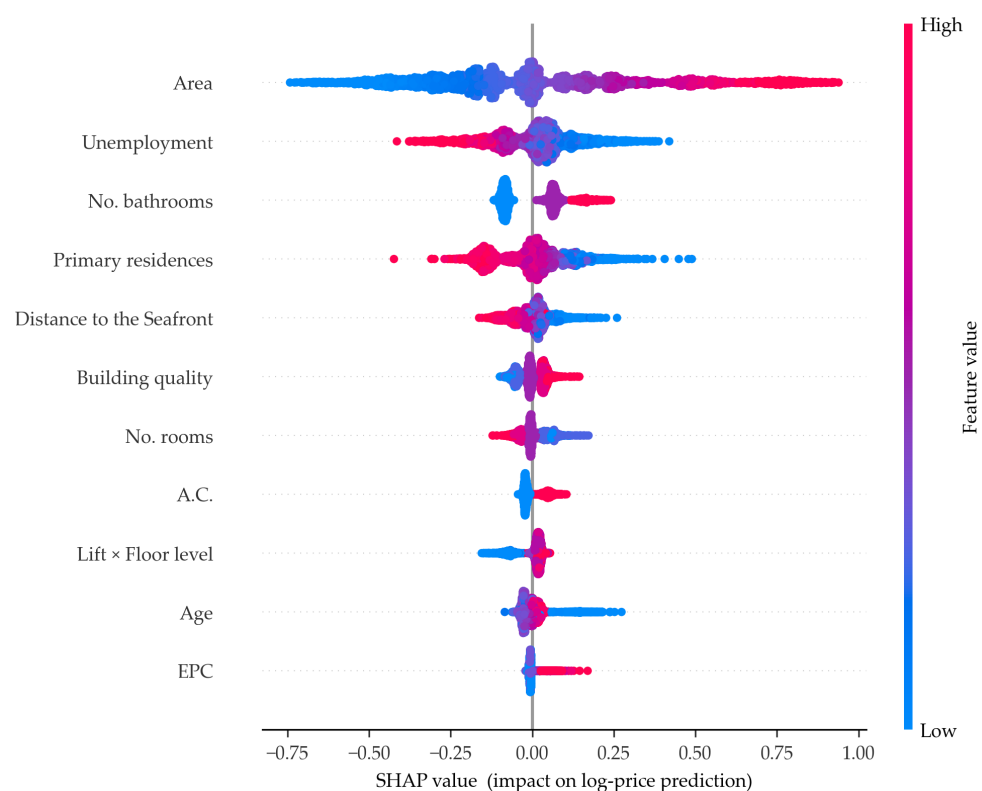
The 70/30 random split yielded a training set of  $n = 4278$  observations and a test set of  $n = 1834$ . The optimized XGBoost model achieves an  $R^2$  of 0.932 on the training set, 0.878 on standard five-fold cross-validation ( $RMSE = 0.199 \pm 0.005$ ), and 0.878 on the test set ( $RMSE = 0.201$ ). The near-perfect alignment between standard cross-validation and holdout performance ( $\Delta R^2 = 0.0002$ ) confirms strong generalization to unseen data with no evidence of systematic overfitting beyond the train/test gap inherent to gradient boosting.

The spatial cross-validation yields a substantially lower  $R^2$  of 0.734 ( $\pm 0.055$ ), a gap of 0.144 relative to standard CV. This divergence can be interpreted as a manifestation of spatial autocorrelation: when entire geographic zones are held out, the model loses access to spatially proximate training observations whose prices would otherwise be closely correlated.

The result indicates that standard cross-validation is optimistic in this spatially structured context, which motivates the use of spatially aware validation as a complementary diagnostic [42].

#### 4.3. SHAP Feature Importance

Figure 1 summarizes the full distribution of SHAP contributions for each feature. Each dot represents one observation; its horizontal position is the SHAP value  $\phi$ , i.e., the signed contribution of that feature to the predicted log-price relative to the model baseline (dots to the right increase the prediction; dots to the left decrease it). Dot color encodes the feature value (low-to-high), so a predominance of warm-colored dots on the right indicates that higher values of the feature tend to raise predicted prices, whereas warm colors on the left indicate the opposite. The vertical spread reflects the density and heterogeneity of effects across observations, and features are ordered by their mean absolute SHAP value (global importance).



**Figure 1.** SHAP values of XGBoost model.

*Area* is the dominant predictor by a wide margin (mean  $|\phi| = 0.262$ ), more than twice the importance of the next feature, consistent with recent explainable ML evidence in hedonic rental price prediction [26]. *Unemployment* (0.092), *No. bathrooms* (0.084), and *Primary residences* (0.083) rank second through fourth, reflecting the combined weight of socioeconomic context and physical quality attributes. *Distance to the Seafront* (0.041) captures Barcelona’s well-established coastal premium, while *Building quality* (0.032), *No. rooms* (0.031), *A.C.* (0.029), *Lift x Floor level* (0.028) and *Age* (0.024) contribute more moderately. *EPC* registers the lowest importance (0.011), consistent with evidence that energy performance ratings command only marginal premiums in Southern European housing markets [4].

Computing the share of total SHAP variance attributable to each feature reveals that *Area* alone accounts for 71.97% of the total variance in SHAP values across the training set, making it the primary driver not only of mean feature importance but also of cross-sectional

variability in SHAP attributions—the dimension directly exploited by the subsequent  $k$ -means clustering step. The implications of this dominance for submarket interpretation are examined in Section 5.

#### 4.4. Identification of Housing Market Segments

$K$ -means clustering on the SHAP value matrix was evaluated for  $k \in \{3, 4, 5\}$ . The silhouette score is highest at  $k = 3$  (0.324), and falls to 0.251 and 0.244 for  $k = 4$  and  $k = 5$ , respectively. The inertia elbow also identifies  $k = 3$  as the point of most rapid diminishing returns. Three clusters were therefore selected.

Cluster labeling was performed post hoc on the basis of average unit price (EUR/m<sup>2</sup>) computed from training-set observations, yielding three market segments: High-End (649 observations; 15.2% of the training set; mean unit price EUR 5502/m<sup>2</sup>; mean total price EUR 930,714), Mid-Range (1796 obs.; 42.0%; EUR 4577/m<sup>2</sup>; EUR 449,938), and Entry-Level (1833 obs.; 42.8%; EUR 3907/m<sup>2</sup>; EUR 249,794). The bimodal distribution, with roughly equal shares in Mid-Range and Entry-Level and a distinct smaller luxury segment, is consistent with the structure of a large southern European metropolitan housing market [4].

#### 4.5. Cluster Stability Analysis

Across  $B = 1000$  spatial block bootstrap iterations, the mean ARI is 0.892 (std = 0.149; range [0.329, 1.000]), well above the 0.80 high-stability threshold [40]. This result confirms that the three-cluster solution is robust to spatial subsampling of the training set: when any one of the five geographic blocks is excluded or replicated, the cluster structure is largely preserved. The non-negligible standard deviation (0.149) reflects the finite-sample variability inherent in block bootstrap with a small number of blocks—in extreme draws where one block dominates, the ARI can fall, though this occurs only rarely (minimum observed ARI = 0.329). Mean cluster-size variability (coefficient of variation = 0.376) is moderate, indicating that segment sizes fluctuate across resamples without dissolving any segment.

#### 4.6. Out-of-Sample Cluster Assignment

SHAP values were recomputed for all 1834 test data observations using the training-set TreeExplainer, and each was assigned to the nearest training centroid in SHAP space. The test data cluster distribution (Mid-Range: 39.7%, Entry-Level: 45.3%, High-End: 15.0%) closely replicates the training-set proportions (42.0%, 42.8%, 15.2%). The differences across segments are small and well within the range attributable to sampling variability, providing evidence that the three market segments generalize to unseen properties without any information leakage from the test set.

#### 4.7. Econometric Validation: OLS Interaction Model

Model 1, estimated on the holdout test set ( $n = 1834$ ) with log-price as the dependent variable, eleven predictors, and two cluster dummies (Entry-Level as reference), achieves  $R^2 = 0.851$  (adjusted  $R^2 = 0.850$ ). The cluster dummies are individually significant— $\hat{\beta}_{\text{Mid-Range}} = 0.120$  ( $p < 0.001$ ),  $\hat{\beta}_{\text{High-End}} = 0.090$  ( $p = 0.015$ )—and the joint Wald test firmly rejects their joint nullity ( $F = 36.39$ ,  $p < 0.001$ ). Cluster membership therefore retains explanatory power for prices after conditioning on all eleven observable property characteristics. This out-of-sample evaluation follows established validation practices in housing submarket research [3]. The result is read in conjunction with the external validation steps that follow, which rely on information entirely independent of the model.

Model 2 augments Model 1 with full interactions between all eleven predictors and the two non-reference cluster dummies (22 interaction terms).  $R^2$  rises to 0.862 ( $\Delta R^2 = 0.011$ ,  $\Delta \text{Adj } R^2 = 0.009$ ), and the interaction block is jointly significant ( $F = 6.41$ ,  $p < 0.001$ ),

indicating heterogeneous attribute semi-elasticities across segments [43]. Table 3 presents the complete coefficient estimates for Model 2. The Net effect (%) column translates each semi-logarithmic coefficient into a percentage change in price, computed as  $(e^{\hat{\beta}} - 1) \times 100$  for main effects and as  $(e^{\hat{\beta}_{\text{main}} + \hat{\delta}} - 1) \times 100$  for interaction rows—where  $\hat{\beta}_{\text{main}}$  is the base coefficient of the variable and  $\hat{\delta}$  is the cluster-specific differential. This allows direct comparison of economic magnitudes across attributes measured on different scales, and makes explicit how much of the estimated return to each characteristic persists within each market tier.

**Table 3.** OLS Model 2 (interaction specification) estimated on the holdout test set.

| Variable                    | Coef.   | Sig. | Net eff. (%) |
|-----------------------------|---------|------|--------------|
| Intercept                   | +13.245 | ***  | —            |
| High-End (intercept shift)  | +0.915  | **   | +149.59      |
| Mid-Range (intercept shift) | +0.445  | †    | +56.09       |
| Area (m <sup>2</sup> )      | +0.011  | ***  | +1.12        |
| × High-End                  | −0.006  | ***  | +0.49        |
| × Mid-Range                 | −0.003  | **   | +0.83        |
| Age (years)                 | +0.000  |      | +0.02        |
| × High-End                  | −0.001  |      | −0.04        |
| × Mid-Range                 | −0.001  | **   | −0.12        |
| Lift × Floor level          | +0.019  | ***  | +1.88        |
| × High-End                  | −0.025  | **   | −0.67        |
| × Mid-Range                 | −0.001  |      | +1.77        |
| Unemployment (%)            | −4.204  | ***  | −98.51       |
| × High-End                  | +0.077  |      | −98.39       |
| × Mid-Range                 | +0.762  |      | −96.80       |
| Building quality            | +0.053  | ***  | +5.42        |
| × High-End                  | −0.018  |      | +3.56        |
| × Mid-Range                 | −0.011  |      | +4.31        |
| EPC                         | +0.015  | **   | +1.50        |
| × High-End                  | −0.002  |      | +1.32        |
| × Mid-Range                 | +0.001  |      | +1.56        |
| Dist. Seafront (m)          | −0.000  | ***  | −0.01        |
| × High-End                  | +0.000  | **   | −0.00        |
| × Mid-Range                 | +0.000  | †    | −0.00        |
| A.C.                        | +0.077  | ***  | +8.04        |
| × High-End                  | −0.001  |      | +7.91        |
| × Mid-Range                 | −0.001  |      | +7.96        |
| No. rooms                   | −0.039  | **   | −3.86        |
| × High-End                  | −0.002  |      | −4.07        |
| × Mid-Range                 | −0.031  | †    | −6.80        |
| No. bathrooms               | +0.148  | ***  | +15.96       |
| × High-End                  | −0.092  | **   | +5.72        |
| × Mid-Range                 | −0.022  |      | +13.40       |
| Primary residences (%)      | −1.619  | ***  | −80.18       |
| × High-End                  | −0.009  |      | −80.35       |
| × Mid-Range                 | −0.078  |      | −81.68       |

\*\*\*  $p < 0.001$ ; \*\*  $p < 0.01$ ; \*  $p < 0.05$ ; †  $p < 0.10$ .

The estimated coefficient on *Area* is significantly lower in Mid-Range ( $\hat{\delta} = -0.003$ ,  $p = 0.006$ ) and lower still in High-End ( $\hat{\delta} = -0.006$ ,  $p < 0.001$ ), relative to Entry-Level. This pattern is consistent with lower model-attributed price differentiation from additional floor area in the High-End segment, where qualitative attributes, finishes, and location may play a comparatively larger role. The *Age* coefficient is steeper in Mid-Range ( $\hat{\delta} = -0.001$ ,  $p = 0.002$ ), consistent with a segment where newer or renovated stock is more strongly associated with predicted prices than in Entry-Level. The *Lift × Floor level* premium is significantly reduced in High-End ( $\hat{\delta} = -0.025$ ,  $p = 0.006$ ), consistent with floor-level accessibility features contributing less to within-segment price differentiation in higher-tier

properties. Finally, *No. bathrooms* generates a substantially smaller estimated coefficient in High-End ( $\hat{\delta} = -0.092$ ,  $p = 0.004$ ); the attenuated coefficient suggests that additional bathrooms contribute less to within-segment price differentiation, possibly because higher bathroom counts are more common in that segment.

These patterns are consistent with submarket theory, under which implicit attribute prices are expected to differ systematically across market segments [1].

Given that Model 2 confirms statistically significant differences between Entry-Level and both higher-tier segments, an alternative specification with Mid-Range as the reference category was estimated to directly characterise the Entry-Level–Mid-Range and Mid-Range–High-End contrasts. Table 4 presents the significant interaction terms ( $p < 0.1$ ).

**Table 4.** OLS Model 3 with Mid-Range as reference category, estimated on the holdout test set.

| Variable                             | Coef.   | Sig. | Net eff. (%) |
|--------------------------------------|---------|------|--------------|
| Intercept                            | +13.690 | ***  | –            |
| <i>Entry-Level</i> (intercept shift) | −0.445  | †    | −35.94       |
| <i>High-End</i> (intercept shift)    | +0.469  |      | +59.90       |
| <i>Area</i> (m <sup>2</sup> )        | +0.008  | ***  | +0.83        |
| × <i>Entry-Level</i>                 | +0.003  | **   | +1.12        |
| × <i>High-End</i>                    | −0.003  | ***  | +0.49        |
| <i>Age</i> (years)                   | −0.001  | ***  | −0.12        |
| × <i>Entry-Level</i>                 | +0.001  | **   | +0.02        |
| <i>Lift × Floor level</i>            | +0.018  | ***  | +1.77        |
| × <i>High-End</i>                    | −0.024  | *    | −0.67        |
| <i>Dist. Seafront</i> (m)            | −0.000  | ***  | −0.00        |
| × <i>Entry-Level</i>                 | −0.000  | †    | −0.01        |
| × <i>High-End</i>                    | +0.000  | †    | −0.00        |
| <i>No. rooms</i>                     | −0.070  | ***  | −6.80        |
| × <i>Entry-Level</i>                 | +0.031  | †    | −3.86        |
| <i>No. bathrooms</i>                 | +0.126  | ***  | +13.40       |
| × <i>High-End</i>                    | −0.070  | *    | +5.72        |

\*\*\*  $p < 0.001$ ; \*\*  $p < 0.01$ ; \*  $p < 0.05$ ; †  $p < 0.10$ .

In this specification, the intercept shift for Entry-Level remains marginally significant ( $\hat{\beta} = -0.445$ ,  $p < 0.10$ ), confirming a structural price gap below Mid-Range that is not fully absorbed by attribute differences. By contrast, the intercept shift for High-End is not statistically distinguishable from Mid-Range ( $\hat{\beta} = +0.469$ ,  $p = 0.275$ ), suggesting that once attribute differences and their differential valuations are accounted for, the baseline price levels of High-End and Mid-Range properties converge.

The area saturation gradient extends across all three tiers: the estimated coefficient on *Area* is higher in Entry-Level than in Mid-Range ( $\hat{\delta} = +0.003$ ,  $p < 0.01$ ) and lower still in High-End ( $\hat{\delta} = -0.003$ ,  $p < 0.001$ ), relative to Mid-Range. *Age* depreciation is steeper in Mid-Range than in Entry-Level ( $\hat{\delta}_{EL} = +0.001$ ,  $p < 0.01$ ), consistent with older stock being more strongly associated with lower predicted prices in the mid-market segment. The *Lift × Floor level* premium and the *No. bathrooms* premium are both significantly attenuated in High-End relative to Mid-Range ( $\hat{\delta} = -0.024$ ,  $p < 0.05$ ; and  $\hat{\delta} = -0.070$ ,  $p < 0.05$ , respectively), consistent with the saturation effects identified in the primary specification.

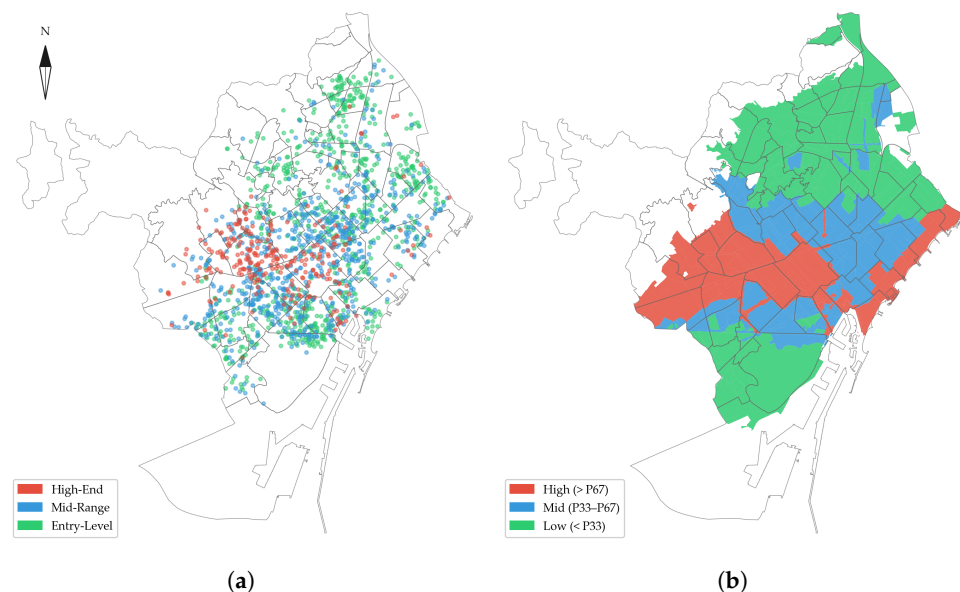
#### 4.8. External Variable Validation

Of the 1834 holdout test observations, 11 fell within single-family cadastral zones and were excluded, and 5 could not be matched to any zone, leaving 1818 observations for this analysis. Shapiro–Wilk tests confirm non-normality in all three clusters ( $p < 0.001$  in all cases), justifying the non-parametric approach.

The Kruskal–Wallis test rejects the null hypothesis of equal cadastral value distributions across clusters with overwhelming significance ( $H = 474.68$ ,  $p < 0.001$ ). Post-hoc

pairwise Mann–Whitney  $U$  tests with Bonferroni correction confirm that all three cluster pairs are mutually distinct (all corrected  $p < 0.001$ ). The gradient is monotonically ordered and consistent with the market-tier labeling derived from SHAP profiles, as shown in Figure 2: High-End properties are located in cadastral zones with the highest mean residential value (EUR 1390/m<sup>2</sup>; median EUR 1354/m<sup>2</sup>), followed by Mid-Range (EUR 1114/m<sup>2</sup>; median EUR 1101/m<sup>2</sup>) and Entry-Level (EUR 861/m<sup>2</sup>; median EUR 859/m<sup>2</sup>).

The monotonic correspondence between SHAP-derived valuation tiers and an independent administrative benchmark of relative residential value provides evidence that the segments align with a spatially structured external indicator, consistent with the model having internalized patterns related to market stratification rather than spuriously recovering observable correlates.



**Figure 2.** (a) Spatial distribution of housing submarket (holdout test set). (b) 2017 cadastral value (administratively defined, market-informed benchmark) per cadastral valuation zone [41], classified into terciles (low, medium, high). Both graphics show neighborhoods across Barcelona.

#### 4.9. Geographic Coherence Validation

Within test data, a chi-squared test of independence between cluster membership and Barcelona neighborhood (65 neighborhoods;  $n = 1834$ ) firmly rejects spatial randomness ( $\chi^2 = 812.71$ ,  $df = 128$ ,  $p < 0.001$ ). The High-End segment concentrates disproportionately in Sant Gervasi–Galvany (22.5% of its holdout observations) and la Dreta de l'Eixample (15.3%), while Entry-Level dominates in more peripheral and socioeconomically diverse areas such as el Raval (5.2%), and la Verneda i la Pau (4.8%). The Mid-Range segment is more evenly distributed across transitional and central-peripheral neighborhoods.

Spatial autocorrelation of cluster membership, quantified via Moran's  $I$  with a  $k = 15$  KNN weight matrix on the ordinal price-tier encoding described in Section 3, is positive and highly significant ( $I = 0.342$ ,  $z = 42.23$ ,  $p < 0.001$ ), confirming that geographically proximate properties tend to share segment membership more often than chance. At the same time, the moderate magnitude indicates that the segmentation is not reducible to a spatial indicator: substantial within-neighborhood heterogeneity in cluster assignment persists, consistent with a valuation-based rather than purely geographic segmentation [17].

## 5. Discussion and Limitations

This analysis demonstrates that ML models can capture systematic heterogeneity patterns that satisfy multiple independent validation criteria. The identified submarkets

exhibit: (1) computational stability under spatial block bootstrap (mean ARI 0.892), (2) generalization to out-of-sample data, (3) coherence with an external indicator never used in training (Kruskal–Wallis  $H = 474.68$ ,  $p < 0.001$ ), (4) statistically significant differential valuation of attributes (interaction  $F = 6.41$ ,  $p < 0.001$ ), and (5) meaningful geographic structure without being reducible to pure location (Moran's  $I = 0.342$ ). The convergence of evidence across these independent tests, each designed to detect different types of artifacts, substantially reduces the probability that findings merely reflect computational quirks.

Two primary concerns require careful interpretation. First, SHAP value reliability: tree-based models may produce unreliable SHAP values in regions of the feature space that were sparsely populated during training, particularly when marginalizing correlated features. This concern is addressed proactively at three levels. At the estimation level, we use path-dependent feature perturbation, which constrains SHAP computation to feature combinations that actually occurred in the training data, avoiding evaluation at empirically unsupported points in the feature space. At the variable preparation level, correlated predictors were collapsed into PCA-derived composite indices during preprocessing, preventing the correlated individual variables from entering the selection pool and potentially distorting the model. Although neither composite index was retained in the final specification—the centrality index was non-significant in the SDM, and the material conditions index exceeded the VIF threshold—this preprocessing step limited the multicollinearity pathways available to the selected variables, all of which have  $VIF < 3$ . Low-prevalence binary variables (*New-build* and *Large terrace*) were additionally excluded, reducing the risk of underrepresented leaf regions. At the validation level, if residual SHAP unreliability still dominated the analysis, we would expect clusters to be unstable under resampling, fail to generalize, and show no coherence with external variables—none of which is observed.

Second, the OLS interaction validation is inherently circular: since clusters derive from SHAP values computed from XGBoost predictions, any systematic pattern the model has learned—whether genuine market heterogeneity or a modeling artifact—will necessarily produce clusters that differ in their price-conditional means. The OLS results therefore confirm that observed patterns are statistically systematic and replicable on held-out data, but cannot resolve whether they reflect economic structure or a pipeline artifact. This inherent limitation is mitigated through two complementary strategies: (1) the OLS is estimated exclusively on a held-out test set not used in training or clustering, so statistical significance reflects out-of-sample generalization rather than in-sample overfitting; and (2) the two subsequent validation steps—external cadastral value coherence and geographic coherence—draw on information entirely independent of the model, breaking the circular dependence on model-derived quantities. Together, these steps support interpreting the overall findings as exploratory evidence of heterogeneity patterns in XGBoost's learned valuation structure that persist under spatial resampling, align with an independent administrative economic indicator, and exhibit non-trivial geographic structure not reducible to location alone.

While no single validation resolves all concerns, their convergence strengthens the overall credibility of the findings: each test addresses a different potential artifact, and the consistency of results across independent checks reduces the likelihood that the identified patterns are spurious.

The dominance of *Area* in SHAP variance (accounting for 71.97% of the total) reflects that the model learned floor area to be the primary price determinant, a finding consistent with recent SHAP-based hedonic ML evidence [26]. However, two lines of evidence suggest that the segments capture more than a pure size typology, albeit modestly.

First, the external cadastral validation uses the *ponencia de valores* expressed per square metre—a benchmark that explicitly standardises by floor area. The strong monotonic alignment between cluster ordering and cadastral value per m<sup>2</sup> ( $H = 474.68$ ,  $p < 0.001$ ; all pairwise comparisons significant after Bonferroni correction) would not be expected if the clusters were merely sorting apartments by size.

Second, the OLS interaction model shows that the statistically significant interaction terms concentrate on physical attributes—*Area*, *Age*, *Lift × Floor level*, *No. bathrooms*, and *Distance to the Seafront*—rather than on the socioeconomic variables. The interaction terms for *Unemployment rate* and *Primary residences share* are not individually significant, indicating that their influence on prices operates with similar magnitude across the three segments. These neighborhood-level variables do, however, retain strong and highly significant main effects in all three tiers, consistent with their role as location proxies that differentiate price levels *between* neighborhoods even though their marginal valuation does not differ *across* segments. It is plausible that these variables partly capture unobserved neighborhood characteristics—amenities, social composition, urban quality—that are not directly measured in the dataset but contribute to the spatial differentiation of housing submarkets.

Taken together, the evidence supports interpreting the three segments primarily as a size–price gradient with secondary differentiation along physical quality and location dimensions. The clusters are better characterised as correlated size–quality–location profiles than as richly multidimensional submarkets, though the cadastral validation and the significant physical-attribute interactions indicate that the segmentation is not reducible to floor area alone.

Specifically, the between-cluster sum of squares attributable to *Area* SHAP alone represents 92.8% of the total between-cluster sum of squares (see Supplementary Materials S4. *Area dominance analysis*), confirming that the primary axis of segmentation is the price-tier/size gradient. The remaining 7.2%—distributed across *Unemployment rate*, *No. bathrooms*, *Primary residences share*, and other variables—is modest in absolute terms. The dominance of *Area* is both a substantive result—reflecting the centrality of floor area in hedonic price formation—and an interpretive constraint on the multidimensional richness of the segmentation. This pattern is not unique to Barcelona: Gyger et al. [26], analyzing over 1.5 million German rental listings with gradient boosting and SHAP, find that usable space is the single most influential predictor by a wide margin, affecting predictions by approximately 25% on average—more than twice the next variable. The consistency of this finding across two distinct European housing markets suggests that the dominance of floor area in SHAP-based decompositions reflects a structural feature of hedonic price formation rather than an artifact of the present dataset or variable specification. Moreover, floor area may partly act as a proxy for socioeconomic and locational factors not fully captured by the available covariates: in Barcelona, larger apartments tend to concentrate in central and affluent neighborhoods, while smaller units are more prevalent in working-class and peripheral areas. To the extent that this spatial sorting is not entirely absorbed by the neighborhood-level variables already included in the model, the SHAP contribution attributed to *Area* may partly reflect these underlying locational and socioeconomic gradients.

A direct sensitivity test confirms this finding: running *k*-means ( $k = 3$ ) on the ten-feature SHAP matrix with *Area* excluded produces a fundamentally different partition (Adjusted Rand Index = 0.08, silhouette = 0.263 versus 0.324), confirming that *Area* SHAP is the primary cluster separator.

A key data limitation concerns the use of listing prices rather than transaction prices. Asking prices reflect seller expectations and strategic pricing behavior, and may deviate from market-clearing equilibrium prices. As introduced in Section 1, this has direct implications for interpreting SHAP contributions as valuations: the model captures the structure

of seller-stated prices rather than equilibrium implicit prices in the sense of Rosen [2]. Nonetheless, the relative ordering of attribute contributions across submarkets—which the clustering and validation procedures exploit—may still be informative about market heterogeneity, since systematic differences in how sellers price attributes across segments are themselves a structural feature of the market. Validation using transaction prices combined with negotiation outcomes, time-on-market, and buyer characteristics would provide a more direct test of whether the identified segmentation aligns with market-clearing behavior. As a reference, an analogous concern has been documented in the German rental market, where listing data from online platforms reflect landlords' pricing intentions rather than amounts ultimately paid by tenants, and the magnitude of the gap between asking and achieved rents varies by market tightness and price segment [26]. Without claiming that this observation generalizes to all European markets, the German case illustrates that the gap between asking and realized prices is a recognized limitation in the literature, reinforcing the interpretive caution adopted here.

A related temporal consideration concerns the data extraction date (17 March 2020). As described in Section 3, the data were collected three days after Spain declared its state of alarm, with 91.1% of listings last modified before the alarm declaration. Although systematic repricing within the first three days of the lockdown is unlikely, this cannot be verified directly with the available data. The results should therefore be interpreted as reflecting a predominantly pre-lockdown asking-price structure observed at the onset of the pandemic. Importantly, the validation tests that rely on the spatial relative ordering of submarkets are robust to monotonic level shifts and would not be materially affected by modest lockdown-induced adjustments to absolute asking prices.

A further temporal gap concerns the external validation, which employs 2017 cadastral values against 2020 listings. No urban development operations or market dynamics of sufficient magnitude to alter the relative ordering between zones occurred during this interval. Moreover, cadastral values are administratively updated at irregular intervals incorporating smoothed market trends, so the spatial relative ordering of cadastral zones—which is what the Kruskal–Wallis test exploits—is substantially more stable than absolute values. The non-parametric rank-based test is robust to monotonic level shifts and thus insensitive to the specific year of the cadastral assessment so long as the relative spatial stratification persists.

## 6. Conclusions

This research demonstrates that ML models like XGBoost can identify systematic heterogeneity patterns in housing markets without requiring predefined spatial boundaries. By combining tree-based modeling with SHAP-based interpretability and multi-pronged validation, we establish a framework for revealing market structures that traditional approaches leave latent. Applied to 6112 Barcelona housing listings (2020), the approach identifies three empirically distinct submarkets—High-End, Mid-Range, Entry-Level—that satisfy rigorous validation criteria: computational stability under spatial resampling (ARI 0.892), out-of-sample generalizability, differential physical-attribute semi-elasticities detected by OLS interaction models, coherence with an independent administrative indicator, and meaningful geographic structure without pure spatial determination.

The segmentation is primarily driven by the size–price gradient, with secondary differentiation along physical quality and location dimensions, suggesting that the identified segments are best interpreted as correlated size–quality–location profiles with modest multidimensional richness.

Unlike prior ML studies focused solely on prediction, we demonstrate that interpretable models can reveal market structure when paired with explicit artifact detection. Unlike traditional submarket studies imposing boundaries *a priori*, we recover segments

endogenously from valuation patterns. This bridges the gap between predictive power and economic interpretability—a frontier in housing market analysis.

These findings carry practical implications that, while grounded in the model’s learned valuation structure rather than causal identification, point to directions worth exploring. A clarification on scope is in order: the contribution is a diagnostic taxonomy of the market with an interpretable valuation profile per segment, not a competitor to automated valuation or mass-appraisal models, which already internalise attribute heterogeneity within their predictive machinery. The instruments where SHAP-based segmentation adds genuine signal are therefore those where a human agent—regulator, planner, or policymaker—requires an interpretable map of the market, rather than those where another predictive model would consume the segmentation as an additional feature.

For financial regulators, an externally trained SHAP-based segmentation could serve as an independent benchmark against which to audit commercial automated valuation models for systematic biases across segments, and could inform segment-aware stress tests of mortgage portfolios and concentration monitoring of bank exposure across submarkets—complements to currently uniform macroprudential frameworks. For urban planners, the segmentation provides a diagnostic layer for ex-ante evaluation of the differential impact of infrastructure investments, rezoning, or regulatory changes across submarkets that would otherwise remain latent within city-level aggregates. For policymakers, the explicit segment labels offer a basis for targeted instruments—subsidies to first-time buyers concentrated in the Entry-Level segment, differentiated property tax brackets, or eligibility criteria for protected housing programmes—where uniform city-wide rules implicitly assume a homogeneity that the validation evidence does not support.

Some limitations warrant attention. Our analysis relies on listing rather than transaction prices, potentially reflecting strategic pricing behavior; validation with transaction prices, time-on-market, and buyer characteristics would strengthen inferences about actual market outcomes. Additionally, confirming findings through alternative methodological pipelines—Bayesian hierarchical models, finite mixture models, or switching regressions—would establish whether heterogeneity detection is specific to XGBoost-SHAP or generalizes across methods. Future research could extend this approach through temporal dynamics studying submarket evolution, spatially constrained clustering explicitly testing trade-offs between functional coherence and geographic contiguity, application to newer tree-based models, and quasi-experimental designs exploiting regulatory or infrastructure shocks to enable causal inference about segmentation mechanisms.

**Supplementary Materials:** The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/urbansci10080440/s1>, Supplementary Materials S1. *Variable calculation methods*; Supplementary Materials S2. *Clustering method sensitivity*; Supplementary Materials S3. *Code notebooks and variable documentation*; Supplementary Materials S4. *Area dominance analysis*.

**Author Contributions:** Conceptualization, C.M.-D.; methodology, J.R.-Q.; investigation, J.R.-Q.; data curation, J.R.-Q.; formal analysis, J.R.-Q.; validation, J.R.-Q. and C.M.-D.; visualization, J.R.-Q.; writing—original draft preparation, J.R.-Q.; writing—review and editing, C.M.-D.; supervision, C.M.-D. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by the AGAUR-FI Joan Oró predoctoral fellowship program (2023-FI-1 00598 grant), from the Office for Universities and Research of the Department of Research and Universities of the Generalitat de Catalunya, and the European Social Fund Plus.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Due to Spanish data protection regulations on sensitive geolocation data, the raw dataset cannot be publicly shared. All analysis code is documented in Jupyter Notebooks provided as Supplementary Materials S3. *Code notebooks and variable documentation.*

**Acknowledgments:** The authors thank Paúl Espinoza-Zambrano for converting the cadastral valuation zones into vector files, which were essential for the spatial analysis conducted in this study. Moreover, we wish to express our gratitude to Adrià Rojas for his invaluable support. This paper is framed in the Doctoral Thesis of the first author and linked to the EnerValor 3 project “Integrating the potential of urban planning, land management, and One-Stop Shops to foster deep retrofits in disadvantaged neighborhoods and specific heritage buildings by leveraging land value gains and comprehensive implementation agencies”. During the preparation of the manuscript, the authors used LLMs for the purposes of proofreading and improving the clarity of the written text, and for drafting sections of the methodology based on review of annotated Jupyter notebooks. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

|         |                                |
|---------|--------------------------------|
| AIC     | Akaike Information Criterion   |
| ARI     | Adjusted Rand Index            |
| CV      | Cross-Validation               |
| EPC     | Energy Performance Certificate |
| IQR     | Interquartile Range            |
| KNN     | <i>k</i> -Nearest Neighbors    |
| ML      | Machine Learning               |
| OLS     | Ordinary Least Squares         |
| OOB     | Out-of-Bag                     |
| PCA     | Principal Component Analysis   |
| RMSE    | Root Mean Squared Error        |
| SDM     | Spatial Durbin Model           |
| SHAP    | Shapley Additive Explanations  |
| VIF     | Variance Inflation Factor      |
| XGBoost | Extreme Gradient Boosting      |

## References

1. Watkins, C.A. The Definition and Identification of Housing Submarkets. *Environ. Plan. A Econ. Space* **2001**, *33*, 2235–2253. [\[CrossRef\]](#)
2. Rosen, S. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *J. Political Econ.* **1974**, *82*, 34–55. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Goodman, A.C.; Thibodeau, T.G. Housing Market Segmentation and Hedonic Prediction Accuracy. *J. Hous. Econ.* **2003**, *12*, 181–201. [\[CrossRef\]](#)
4. Marmolejo-Duarte, C.; Chen, A. The Uneven Price Impact of Energy Efficiency Ratings on Housing Segments. Implications for Public Policy and Private Markets. *Sustainability* **2019**, *11*, 372. [\[CrossRef\]](#)
5. Royuela, V.; Duque, J.C. HouSI: Heuristic for Delimitation of Housing Submarkets and Price Homogeneous Areas. *Comput. Environ. Urban Syst.* **2013**, *37*, 59–69. [\[CrossRef\]](#)
6. Rico-Juan, J.R.; Taltavull de La Paz, P. Machine Learning with Explainability or Spatial Hedonics Tools? An Analysis of the Asking Prices in the Housing Market in Alicante, Spain. *Expert Syst. Appl.* **2021**, *171*, 114590. [\[CrossRef\]](#)
7. Valier, A. Who Performs Better? AVMs vs Hedonic Models. *J. Prop. Invest. Financ.* **2020**, *38*, 213–225. [\[CrossRef\]](#)
8. Jha, S.B.; Pandey, V.; Jha, R.K.; Babiceanu, R.F. Machine Learning Approaches to Real Estate Market Prediction Problem: A Case Study. *arXiv* **2020**, arXiv:2008.09922. [\[CrossRef\]](#)
9. Sharma, H.; Harsora, H.; Ogunleye, B. An Optimal House Price Prediction Algorithm: XGBoost. *Analytics* **2024**, *3*, 30–45. [\[CrossRef\]](#)

10. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
11. Ardila, D.; Ahmed, A.; Sornette, D. Comparing ask and transaction prices in the Swiss housing market. *Quant. Financ. Econ.* **2021**, *5*, 67–93. [[CrossRef](#)]
12. Lancaster, K.J. A New Approach to Consumer Theory. *J. Political Econ.* **1966**, *74*, 132–157. [[CrossRef](#)] [[PubMed](#)]
13. Anselin, L. *Exploring Spatial Data with GeoDa: A Workbook*; Center for Spatially Integrated Social Science: Santa Barbara, CA, USA, 2005.
14. LeSage, J.P. What Regional Scientists Need to Know about Spatial Econometrics. *Rev. Reg. Stud.* **2014**, *44*, 13–32. [[CrossRef](#)]
15. Brunsdon, C.; Fotheringham, S.; Charlton, M. Geographically Weighted Regression-Modelling Spatial Non-Stationarity. *J. R. Stat. Soc. Ser. D (Stat.)* **1998**, *47*, 431–443. [[CrossRef](#)]
16. Zietz, J.; Zietz, E.N.; Sirmans, G.S. Determinants of House Prices: A Quantile Regression Approach. *J. Real Estate Financ. Econ.* **2008**, *37*, 317–333. [[CrossRef](#)]
17. Bourassa, S.C.; Hamelink, F.; Hoesli, M.; MacGregor, B.D. Defining Housing Submarkets. *J. Hous. Econ.* **1999**, *8*, 160–183. [[CrossRef](#)]
18. Roca Cladera, J. *La Estructura de Valores Urbanos: Un Análisis Teórico-Empírico*; Instituto de Estudios de Administración Local: Madrid, Spain, 1988.
19. Wu, C.; Sharma, R. Housing Submarket Classification: The Role of Spatial Contiguity. *Appl. Geogr.* **2012**, *32*, 746–756. [[CrossRef](#)]
20. Biswas, A.S. Housing Submarkets and the Impacts of Foreclosures on Property Prices. *J. Hous. Econ.* **2012**, *21*, 235–245. [[CrossRef](#)]
21. Bourassa, S.C.; Hoesli, M.; Peng, V.S. Do Housing Submarkets Really Matter? *J. Hous. Econ.* **2003**, *12*, 12–28. [[CrossRef](#)]
22. Tapia, J.; Chavez-Garzon, N.; Pezoa, R.; Suarez-Aldunate, P.; Pilleux, M. Comparing Automated Valuation Models for Real Estate Assessment in the Santiago Metropolitan Region: A Study on Machine Learning Algorithms and Hedonic Pricing with Spatial Adjustments. *PLoS ONE* **2025**, *20*, e0318701. [[CrossRef](#)] [[PubMed](#)]
23. Yoshida, T.; Murakami, D.; Seya, H. Spatial Prediction of Apartment Rent Using Regression-Based and Machine Learning-Based Approaches with a Large Dataset. *J. Real Estate Financ. Econ.* **2024**, *69*, 1–28. [[CrossRef](#)]
24. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [[CrossRef](#)]
25. Hong, J.; Choi, H.; Kim, W.s. A house price valuation based on the random forest approach: The mass appraisal of residential property in South Korea. *Int. J. Strateg. Prop. Manag.* **2020**. [[CrossRef](#)]
26. Gyger, T.; Hauri, S.; Bühlmann, S.; Lehner, M.; Schlesinger, J.; Sigrist, F. Explainable spatial machine learning for hedonic real estate modeling. *Real Estate Econ.* **2026**, *54*, 762–807. [[CrossRef](#)]
27. Tchuente, D. Real Estate Automated Valuation Model with Explainable Artificial Intelligence Based on Shapley Values. *J. Real Estate Financ. Econ.* **2024**. [[CrossRef](#)]
28. Li, Z. Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. *Comput. Environ. Urban Syst.* **2022**, *96*, 101845. [[CrossRef](#)]
29. Spanish Statistical Office. Population and Housing Census. 2021. Available online: <https://www.ine.es/censos2021/> (accessed on 10 May 2025).
30. Spanish Statistical Office. Household Income Distribution Atlas. Available online: [https://www.ine.es/en/experimental/atlas/experimental\\_atlas\\_en.htm](https://www.ine.es/en/experimental/atlas/experimental_atlas_en.htm) (accessed on 10 May 2025).
31. OpenStreetMap Contributors. OpenStreetMap. Available online: <https://www.openstreetmap.org/copyright> (accessed on 1 May 2024).
32. Hooker, G.; Mentch, L.; Zhou, S. Unrestricted Permutation forces Extrapolation: Variable Importance Requires at least One More Model, or There Is No Free Variable Importance. *arXiv* **2021**, arXiv:1905.03151. [[CrossRef](#)]
33. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Anchorage, AK, USA, 4–8 August 2019; pp. 2623–2631. [[CrossRef](#)]
34. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [[CrossRef](#)] [[PubMed](#)]
35. Chen, H.; Covert, I.C.; Lundberg, S.M.; Lee, S.I. Algorithms to estimate Shapley value feature attributions. *Nat. Mach. Intell.* **2023**, *5*, 590–601. [[CrossRef](#)]
36. Shao, X. The Dependent Wild Bootstrap. *J. Am. Stat. Assoc.* **2010**, *105*, 218–235. [[CrossRef](#)]
37. Kurisu, D.; Kato, K.; Shao, X. Gaussian Approximation and Spatially Dependent Wild Bootstrap for High-Dimensional Spatial Data. *J. Am. Stat. Assoc.* **2024**, *119*, 1820–1832. [[CrossRef](#)]
38. Masoero, L.; Thomas, E.; Parmigiani, G.; Tyekucheva, S.; Trippa, L. Cross-Study Replicability in Cluster Analysis. *Stat. Sci.* **2023**, *38*, 105–125. [[CrossRef](#)] [[PubMed](#)]

39. Lubbe, S. Bootstrapping Cluster Analysis Solutions with the R Package ClusBoot. *Austrian J. Stat.* **2024**, *53*, 1–19.
40. Yu, H.; Chapman, B.; Di Florio, A.; Eischen, E.; Gotz, D.; Jacob, M.; Blair, R.H. Bootstrapping estimates of stability for clusters, observations and model selection. *Comput. Stat.* **2019**, *34*, 349–372. [[CrossRef](#)]
41. Dirección General del Catastro. Ponencia de Valores. Available online: [https://www.catastro.hacienda.gob.es/es-ES/ponencia\\_valores.html](https://www.catastro.hacienda.gob.es/es-ES/ponencia_valores.html) (accessed on 6 March 2026).
42. Deppner, J.; Cajias, M. Accounting for Spatial Autocorrelation in Algorithm-Driven Hedonic Models: A Spatial Cross-Validation Approach. *J. Real Estate Financ. Econ.* **2024**, *68*, 235–273. [[CrossRef](#)]
43. Tu, Y.; Sun, H.; Yu, S.M. Spatial Autocorrelations and Urban Housing Market Segmentation. *J. Real Estate Financ. Econ.* **2007**, *34*, 385–406. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.