ISSN: 1648-715X / eISSN: 1648-9179

VILNIUS TECH
Vilnius Gediminas Technical University

INTERNATIONAL JOURNAL OF STRATEGIC PROPERTY MANAGEMENT

2026
Volume 30
Issue 3
Pages 259–272
https://doi.org/10.3846/ijspm.2026.27872

OPPORTUNITIES FOR INVESTMENT IN AIRBNB SHORT-TERM RENTAL PROPERTIES: PREDICTIVE MODELLING FOR CATEGORISATION AND REVENUE ESTIMATION

Natalie NASH, Emad OTRI *, Nikolai MOURAVIEV

Dundee Business School, Abertay University, Dundee, UK

Article History:
- received 13 August 2025
- accepted 23 May 2026

Abstract. This paper discusses the rapidly developing property sector in Edinburgh, Scotland, which in recent years has been shaped by growth in tourism and the expansion of the short-term rental market. The study aims to investigate how Airbnb listings can be categorised into the tiers offering low, medium and high potential earnings for investors. Research methodology includes the K-Means clustering, based on features, such as location, availability, property type, number of bedrooms, bathrooms and price. The paper further examines what variables influence the investors' revenue streams most. The application of the Random Forest Model to the K-Means clustering results showed that pricing, availability and guest capacity were the principal variables affecting earnings. Incorporating predictive modelling into the market categorisation enabled this study to offer a practical, data-based framework for potential investors who are considering various options for investment in properties with an intent to offer them for short-term rent via Airbnb.

Keywords: Airbnb, short-term rentals, earnings potential, investment, predictive modelling.

*Corresponding author. E-mail: *e.otri@abertay.ac.uk*

1. Introduction

The exponential growth of Airbnb has fundamentally reshaped the global accommodation industry (Li & Srinivasan, 2019). Using its peer-to-peer model within the shared economy, Airbnb is now a major investment sector with attractive revenue potential (Oskam, 2019). Whilst there are quite a few publications about the company, it is important to explain how this paper is positioned against the full body of Airbnb and short-term rental literature.

The literature's dominant strand focuses on Airbnb's impact on the housing market. Research by Rossi and d'Addona (2025), Congiu et al. (2024), Duso et al. (2024), Bibler et al. (2021), Koster et al. (2021), García-López et al. (2020) and Barron et al. (2020) investigates how Airbnb raises rents, displaces residents and shrinks long-term housing supply, addressing audiences of policymakers, urban economists and housing advocates. The current paper adopts a different perspective: it accepts Airbnb's presence as a market reality and treats investor decision-making as the object of study, rather than a cause of harm to be regulated.

Another strand focuses on regulation. Papers by Gauß et al. (2024), Jin et al. (2024), Bibler et al. (2021) and Nieuwland and van Melik (2020) treat regulation as a variable, comparing regulatory approaches or measuring causal effects on market outcomes. This paper's approach is different: it treats Scotland's 2023 licensing framework – the Licensing Order, the Control Area Regulations, the tiered licensing fees – not as a policy intervention to be evaluated, but as a risk parameter that differentiates investment tiers. As the paper shows, one property cluster carries the highest earnings potential but also the highest regulatory burden. The regulatory literature is silent about this integration of compliance costs into investment decisions; this paper embeds regulation into investment analysis rather than studying it as an outcome.

Yet another strand is the pricing and revenue prediction literature. Papers such as the Athens GWR study (Iliopoulou et al., 2025), Camatti et al. (2024), Cervera et al. (2024) and Tang et al. (2024) apply machine learning (ML) to Airbnb data to predict nightly price, with existing hosts as the intended beneficiaries. This paper diverges in two respects. First, it focuses not on nightly price but on Gross Yearly Estimate – realistic annual revenue based on occupancy-adjusted daily earnings – which is the figure an investor actually needs to evaluate return on capital. Second, the beneficiary is a prospective investor, not a current host. No paper in the pricing prediction literature makes this shift from operational optimisation to pre-acquisition investment appraisal.

In summary, the paper's position in the Airbnb literature is unique along three dimensions: (a) it serves prospective property investors making pre-acquisition decisions, not policymakers, urban economists or existing hosts; (b) it produces a tiered investment framework with revenue estimates, risk profiles and regulatory obligations attached to each tier; and (c) it applies a sequential K-Means clustering and Random Forest pipeline to constructed revenue metrics (Gross Yearly Estimate) from Inside Airbnb data within Edinburgh's legally distinctive regulatory environment – neither a measure of externalities nor a standard ML price prediction exercise.

Edinburgh Council (2023) linked Airbnb growth to rising prices, tourist congestion and rental shortages. Despite the regulations that were introduced by the Scottish Government in 2020 investors continue to buy properties and rent them out via Airbnb. Airbnb has shifted from its original peer-to-peer model to being a strategic investment opportunity and dominating the shared accommodation economy.

Given these trends, this study's motivation is to provide prospective Airbnb hosts in Edinburgh with a data-driven investment framework, the absence of which is a research gap.

By leveraging a Random Forest Machine Learning Model and K-Means Clustering to identify optimal property types or locations for various investment tiers, the paper's overarching objective is to provide actionable insights that would enable prospective hosts to make strategic, risk-adjusted investment decisions based on revenue potential and property market categorisation.

This study intends to answer two interlinked research questions:

RQ1: How could properties be categorised into investment tiers based on a range of factors, such as location, availability, property type and price?

RQ2: What opportunities for revenue generation could properties in various categories, used for short-term rent, offer to investors?

The paper intends to contribute to the Airbnb literature by combining investor orientation, pre-acquisition perceptions, market-tier segmentation, annual revenue and regulatory risk assessment into a single analytical framework that underpins property investment decisions.

2. Literature review

This literature review aims to identify research gaps at the intersection of short-term rental markets and machine learning-based prediction models. The review shows gaps in three areas: in relation to the target variable; regarding how heterogeneity within the short-term rental market is handled; and in translating predictive output into actionable decision-making framework for investors.

2.1. From nightly price to annual investment return: A gap in the target variable

The application of machine learning to short-term rental platforms has grown substantially, yet the overwhelming majority of studies share a common assumption: that the most meaningful predictive output is the nightly listing price. Tang et al. (2024), applying ten machine learning algorithms to Sydney Airbnb data, identify the CatBoost-Regressor as the best-performing model, with nightly price as the dependent variable and implications directed at platform operators and hosts seeking competitive pricing signals. Camatti et al. (2024), using Dutch listing data with financial history variables, similarly target nightly price, framing their contribution as an improvement over traditional linear pricing for host-level optimisation. Cervera et al. (2024) introduce seasonal pricing variation and gastronomy proximity as novel predictors but retain nightly price as the dependent variable, with outputs directed at hosts adjusting their rates across tourism seasons.

This convergence on nightly price reflects the research community's focus on the existing host rather than the prospective investor. For an investor evaluating whether to acquire a property, the nightly rate is at best an intermediate variable. What matters is the annual revenue stream expected under realistic occupancy conditions, adjusted for seasonal demand and minimum stay requirements. Gross yearly revenue, derived from nightly price, occupancy rate and realistic availability, determines whether acquisition costs can be justified and what return on investment can be achieved over a holding period. The reviewed Airbnb-specific pricing studies do not construct or predict this metric. In the broader real estate predictive modelling literature, the focus similarly remains on transaction price – appropriate for mortgage lending and portfolio revaluation, but misaligned with the income-generation logic of short-term rental investment.

The spatial econometric studies, the Athens GWR study (Iliopoulou et al., 2025), the multiscale GWR study of San Francisco (Hur et al., 2024) and the GNNWR study of Shenzhen (Wang et al., 2022), introduce geographic nuance into price prediction but do not reconceptualise the prediction target. Their contributions lie in modelling spatial non-stationarity in the predictors-price relationship. Similarly, the Vienna XGBoost study (Sharma et al., 2024) and the Warsaw study (Jaroszewicz & Horynek, 2024) achieve high accuracy for transaction prices but offer no mechanism for translating outputs into the income-based valuation logic that short-term rental investment requires.

2.2. Market segmentation as a preprocessing step: Structural heterogeneity and its consequences for prediction

A second gap concerns how the predictive modelling literature handles heterogeneity within the short-term rental market. The dominant methodological response has been spatial: GWR, MGWR and GNNWR models allow the relationship between predictors and the target variable to vary across geographic space. The MGWR study of Wuhan (Liu et al., 2022) shows that different predictors operate at different spatial scales and that this structure changes over time. The Athens GWR study (Iliopoulou et al., 2025) similarly finds that pricing effects of capacity, host attributes and review characteristics vary substantially across the city. These are tangible methodological contributions.

However, spatial heterogeneity is only one dimension of market diversity. The four shades study (Lee & Kim, 2023) distinguishes Airbnb listings by type and host profile, finding that entire-home listings operated by multi-unit commercial hosts have fundamentally different housing market impacts than single-listing casual hosts operating private rooms – implying structural, not merely geographic, differences between listings. Barron et al. (2020) similarly distinguish between vacation homes never part of the long-term rental stock and properties that actively switch between markets, arguing this structural distinction drives the adverse housing effects attributed to Airbnb.

Despite these insights, the predictive modelling literature has not translated structural market heterogeneity into a segmentation-first modelling strategy. Camatti et al. (2024) and Tang et al. (2024) apply their models to full datasets without pre-segmenting the market, expecting the algorithm to learn structural distinctions internally. The consequence is that predicted values and feature importance rankings average across structurally distinct segments a premium entire-home listing in a tourist-saturated central neighbourhood and a part-time private room in a residential suburb are treated as observations from the same data-generating process.

The study of 1.35 million Airbnb reviews across four US cities (Guttentag et al., 2026) demonstrates empirically that guest experience profiles differ systematically and substantially by listing type, suggesting that private room listings and entire-home listings attract not merely different prices but different guest demographics, different evaluation criteria, and different satisfaction drivers. If the user base and experience profile differ so fundamentally between listing types, it is reasonable to expect that the revenue dynamics and the variables driving revenue also differ in ways that aggregate models cannot fully capture. The unsupervised clustering approach – using K-Means or similar methods to identify market segments prior to prediction – represents a methodological response to this challenge that has been applied in the broader real estate investment literature but has not been systematically adopted in the Airbnb-specific modelling literature.

2.3. Translating predictive output into actionable investment guidance

The third gap is perhaps the most consequential from a practical standpoint. The predictive modelling literature has made substantial progress in technical accuracy: the Vienna XGBoost study (Sharma et al., 2024) establishes that human-machine hybrid approaches outperform both unaided experts and standalone algorithms, while the Sydney convolutional neural network study (Lee et al., 2024) reduces MAPE to 8.71% by treating geographic location as an image-like spatial input.

What is notably absent is the translation of these technical advances into structured, investor-facing decision frameworks. The reviewed papers consistently conclude by reporting performance metrics and identifying important predictive features, but stop short of organising outputs into investment tiers that a non-specialist investor can operationalise. Tang et al. (2024) note that findings on guest capacity, bedroom count and cancellation policy could inform host strategy, but do not aggregate listings into investment tiers. Even Camatti et al. (2024), who come closest by incorporating financial history variables, frame their contribution as improved pricing intelligence for platform participants rather than an investment appraisal tool.

This gap reflects a broader limitation identified by multiple review papers. Ding et al. (2023) note that Airbnb scholarship clusters around operational practices and external impacts, with limited attention to the investment decision process. Mathotaarachchi et al. (2024) identify limited exploration of hybrid ML frameworks and persistent interpretability challenges, but do not extend this to how predictive outputs should be structured for different end users. Rodriguez-Serrano (2025) addresses interpretability for real estate valuation, but targets professional appraisers and lending institutions, not individual property investors.

The absence of a useful predictive framework for investors constitutes a clear and practically significant research gap that the current study aims to address by going beyond producing an accurate revenue estimate and proceeding to classify that estimate within a structured investment landscape.

The next section outlines the study's methodology, explaining how revenue is estimated and properties categorised. After that, the findings are presented by showing the property clusters. Discussion provides further details. This is followed by a conclusion outlining implications for investors, the study limitations and suggestions for further research.

3. Methodology

3.1. Data description

This study uses the Inside Airbnb (2024) release for Edinburgh, covering listings up to September 2023. After consolidating the source files and applying the preprocessing steps described below, the analytic dataset, as presented in Table 1, comprises 3,793 listings across 111 neighbourhoods. Key variables include nightly price, availability (0–365 days), bedrooms, bathrooms, accommodates, review activity, review scores and revenue indicators derived for analysis. Prices range from £21 to £357 with a median of £149 and mean of £153.78, availability is heterogeneous with a median of 100 days per year. Review scores cluster at the high end (median 4.84), whereas review counts vary widely. Estimated yearly revenue (availability adjusted) spans £45 to £80,069 with a median of £13,959. No missing values were detected in the variables used for modelling. These figures are computed directly from the final analytical file used in the study.

Table 1. Descriptive statistics for key variables in the Edinburgh Airbnb dataset (N = 3,793)

Variable	Count	Mean	Std	Min	Median	Max
price	3793	153.8	69.9	21.0	149.0	357.0
availability_365	3793	134.3	108.4	1.0	100.0	365.0
bedrooms	3793	1.5	0.7	0.0	1.0	6.0
bathrooms	3793	1.2	0.4	0.0	1.0	4.5
accommodates	3793	3.1	1.6	1.0	2.0	12.0
review_scores_rating	3793	4.8	0.3	1.0	4.8	5.0
number_of_reviews	3793	116	143	1	62	1123
number_of_reviews_ltm	3793	31.3	27.3	0.0	24.0	236.0
reviews_per_month	3793	2.8	2.2	0.0	2.3	13.6
revenue_per_day	3793	153.8	69.9	21.0	149.0	357.0
gross_yearly_estimate	3793	20801	19644	45	13959	80069
gross_yearly_full_occupancy	3793	56129	25519	7665	54385	130305
gross_yearly_availability_adjusted	3793	20801	19644	45	13959	80069
gross_yearly_reviews_adjusted	3793	5037	4468	6	3825	34582

The market composition is consistent with a tourism driven city: entire homes/apartments dominate (2,576 listings), followed by private rooms (1,201), with hotel and shared rooms comprising a small minority. Median nightly prices reflect this structure, at £175 for entire homes/apartments, £99 for hotel rooms, £84 for private rooms and £30 for shared rooms.

All listings are anonymised in the public data; no personal information was accessed or processed.

3.2. Data preprocessing

All preprocessing steps were conducted to ensure the dataset reflected realistic investment conditions and was suitable for distance-based modelling. As presented in Figure 1, variables irrelevant to investment focused analysis, (such as scrape, ID, host name), were removed, qualitative review text was excluded, and multiple source files were merged into a single analytical dataset in which duplicates, inconsistent formats and unrealistic entries were corrected. Extreme values, such as nightly rates above £1,000 or minimum stays longer than thirty nights, were excluded to preserve market realism. Categorical variables were prepared for compatibility with Euclidean distance by encoding neighbourhood identifiers as numeric labels that preserved geographic differentiation without imposing ordinality, and by converting room and property types into dummy variables. Continuous features used in k-means clustering, including bedrooms, bathrooms, accommodates, price, availability, revenue indicators and neighbourhood encodings, were standardised to zero mean and unit variance to prevent variables with larger numeric ranges from dominating the distance computation. Equal weighting was applied across all inputs to maintain interpretability and avoid subjective prioritisation. The resulting feature matrix, drawn directly from the cleaned analytical file (N = 3,793; 111 neighbourhoods), combined the core Airbnb attributes (price, availability, bedrooms, bathrooms, accommodates, review activity and location) with derived revenue variables used downstream. Stable optimisation was ensured through k-means++ initialisation with multiple random starts and a fixed seed, and all transformations were validated by examining summary statistics and distributions in the processed dataset. The variables were calculated or assigned as described in Table 2.

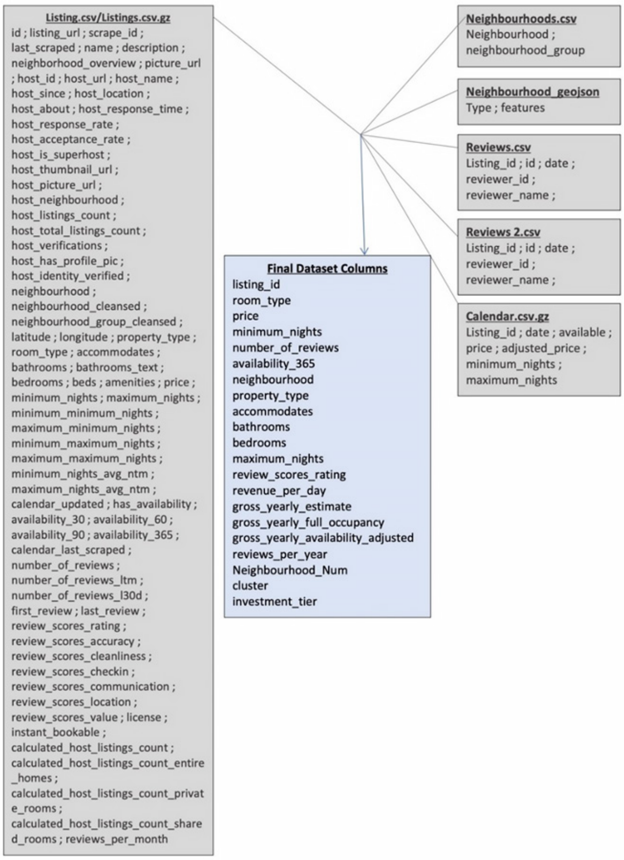

Figure 1. Data integration and transformation workflow for Airbnb analysis

Table 2. Calculated and assigned variables supporting revenue analysis and investment decisions

Metric	Expression	Purpose
Revenue per day	$\textit{Revenue per day} = \textit{Ave price x Occupancy rate}$	Accounting for both pricing and booking frequency, revenue per day provides a baseline measure for use in income generating performance assessment
Gross yearly estimate	$\textit{Gross yearly estimate} = \textit{Revenue per day x} 365$	This metric shows a realistic yearly revenue based on the revenue per day calculation, which accounts for actual booking patterns, derived from occupancy rates. The values assume that the listing continues to generate income at the same price per night rate throughout the year
Reviews per year	$\textit{Reviews per year} = \frac{\textit{Total number of reviews}}{\textit{Years active}}$	This metric divides the review count by the property's active lifespan, providing an indication of a listing's popularity or guest engagement
Gross yearly full occupancy	$\textit{Gross yearly full occupancy} = \textit{Price x} 365$	This value represents the theoretical maximum revenue for a listing, assuming it was booked every day for 365 days. This metric serves as a benchmark for understanding the upper limit of a property's earning potential, under ideal conditions
Gross yearly availability adjusted	$\textit{Gross yearly availability adjusted} = \textit{Revenue per day x Availability} 365$	Derived from adjusting the gross yearly estimate by the number of days a property is available for booking, makes the revenue calculation more realistic. This metric accounts for any factors limiting availability, such as seasonal closures or owner use. ▪ *Gross yearly estimate* predicts actual revenue and compares properties under realistic conditions ▪ *Gross yearly full occupancy* provides an upper limit to assess property's potential revenue

3.3. Analytical and modelling techniques

The analysis follows a Data Driven Decision Making (DDDM) cycle in which data are consolidated, quality checked and transformed into features, exploratory analysis surfaces structure, K means clustering organises listings into interpretable market segments, and supervised models translate those segments into revenue estimates suitable for investment decision making. This sequential design emphasises transparency, reproducibility and actionability.

K means is applied as a segmentation step prior to prediction, not as a mere preprocessing artefact nor as an outcome of the predictive stage. Clustering is performed on bedrooms, bathrooms, accommodates, availability, location encodings and revenue indicators (all standardised), yielding three segments that reflect coherent investment profiles in the Edinburgh market. Multiple values of k were explored; $k = 3$ was selected as the most appropriate balance of parsimony, stability and predictive utility, with additional clusters adding complexity without improving performance. The clusters are subsequently interpreted as lower, medium and higher investment tiers. This outcome is consistent with established clustering practice in real estate and property investment research, where models commonly employ a small number of clusters to maintain conceptual clarity and practical relevance (e.g. Gabriel et al., 2015; Serrano et al., 2020).

The integration of methods is explicit. First, clustering identifies the market structure and assigns each listing to a segment. Second, the resulting cluster label enters the supervised models as an explanatory feature, allowing Linear Regression and Random Forest to produce context aware revenue estimates within and across segments. In other words, segmentation provides interpretable archetypes of the market, and the predictive models quantify expected returns conditional on those archetypes. Empirically, this sequence improved model stability and error relative to unsegmented baselines, which supports the use of segmentation for decision-oriented predictions.

3.4. Robustness: Clustering without revenue-related inputs

To address the potential concern that including revenue signals in clustering could blur the distinction between segmentation and outcome, an additional specification excludes all revenue related variables from the K means stage and clusters only on structural characteristics, availability and location. The resulting labels are then used in the same predictive pipeline and compared with the main specification in terms of segment interpretability and predictive performance. This sensitivity check clarifies the incremental informational value of revenue in segmentation relative to a purely characteristics-based partition and confirms that clustering adds analytical value beyond direct revenue thresholds.

Originally, this study intended to classify properties based on occupancy rates and predict occupancy outcomes. While working with the available data, it became evident that this would not be achievable. Instead, the focus shifted to a revenue- and location-based approach. The new approach used K-Means clustering within the random forest and linear regression models to create an investment tier system. The models provided a more practical framework for the analysis, which allows to assess investment opportunities.

3.5. Practical cluster assignment for new properties

To address the practical case where new listings have no prior revenue data, the clustering workflow incorporates a revenue free assignment step. While revenue variables were included in the explanatory clustering applied to the historical dataset, new properties can be assigned to clusters using only pre-listing characteristics such as bedrooms, bathrooms, accommodates, neighbourhood, expected availability and listed price. We therefore estimate an alternative clustering specification that excludes all revenue related variables and use this model for prospective assignment. This ensures that the framework remains operationally applicable: new properties can be immediately allocated to a segment, after which the predictive models estimate expected revenue conditional on that segment. The revenue free clustering yields a similar structure to the full specification, indicating that segmentation is driven predominantly by property and operational features rather than revenue alone. This modification resolves the cold start limitation identified by the reviewer and ensures the model can be used for genuine forward-looking prediction.

4. Findings

This section presents the findings from the analysis of Edinburgh's Airbnb market, focusing on clustering results, revenue estimations and trends. Clusters were successfully used to categorise properties into investment tiers through K-Means clustering. Key metrics, such as revenue, availability, neighbourhood, bedrooms and bathrooms, were evaluated. Actionable insights for prospective Airbnb hosts have been derived from evaluating these key metrics. Additionally, regression models and exploratory data analysis (EDA) provide further insights into revenue potential within the Edinburgh Airbnb market.

The initial application of the linear regression model to the entire dataset, before clustering, showed a poor prediction performance, with R^2 of 0.72 and RSME (Root squared mean error) of £5,800. Although an R^2 of 0.72 is not unacceptable, it is relatively low, indicating only 72% of the variances in gross yearly estimates were explained, reflecting limited model adequacy. According to Montgomery et al. (2012), the high RSME also indicates significant prediction errors. The variation in estimated yearly gross revenue across clusters is too large for an RSME of £5,800 to be considered acceptable overall, even if this level of error may be acceptable within individual clusters. This outcome suggests that heterogeneity in property characteristics weakened the model's predictive accuracy.

Montgomery et al. (2012) emphasise the importance of model adequacy checking as a critical step in assessing both model fit and appropriateness. Linear regression models assume that residuals are normally distributed, but as illustrated in Figure 2, the skewness and kurtosis indicate deviations from normality. Following the criteria outlined by Lumley et al. (2002) and Chen (2021), high kurtosis suggests disproportionate values (extreme outliers), which can result in an inflated RSME metric.

The linear regression results highlighted that a modified approach was needed to manage the variability in the dataset. To improve predictive accuracy, the data was categorised into three clusters using K-Means clustering, which grouped properties with similar characteristics. When applied to the whole dataset, the Random Forest Model (RFM) achieved a superior performance compared to the linear regression model. As discussed by Hong and Lynn (2020), the RFM model effectively handled the non-linear data better with R^2 = 0.9994 and RSME = 470. Both the RFM and linear regression models were applied to each cluster to evaluate predictive performance. The target value in the analysis is the gross yearly estimate (GYE); the RSME for each model represents monetary differences, and Figure 3 illustrates the comparative performance of the models.

The Random Forest (RF) model was evaluated using a train–test split, with 80% of the data used for training and 20% held out for testing. Model performance was assessed on the test set to reduce overfitting risk. While the RF achieved a high coefficient of determination, this result reflects the model's ability to capture non-linear relationships in the data rather than a guarantee of generalizability. Accordingly, the RF model is used primarily for explanatory and comparative purposes rather than precise prediction.

```
Cluster 0 - Skewness:
price                  : 0.18
availability_365       : -0.16
bedrooms               : 1.14
bathrooms              : 2.17
gross_yearly_estimate  : 0.23

Cluster 0 - Kurtosis:
price                  : -0.49
availability_365       : -1.10
bedrooms               : 1.95
bathrooms              : 6.76
gross_yearly_estimate  : -0.84
```

```
Cluster 1 - Skewness:
price                  : 0.71
availability_365       : 0.53
bedrooms               : 0.41
bathrooms              : 1.38
gross_yearly_estimate  : 1.02

Cluster 1 - Kurtosis:
price                  : -0.58
availability_365       : -0.74
bedrooms               : 0.60
bathrooms              : 2.44
gross_yearly_estimate  : 1.38
```

```
Cluster 2 - Skewness:
price                  : -0.03
availability_365       : 0.69
bedrooms               : 2.72
bathrooms              : 2.45
gross_yearly_estimate  : 1.02

Cluster 2 - Kurtosis:
price                  : -0.96
availability_365       : -0.54
bedrooms               : 12.84
bathrooms              : 9.95
gross_yearly_estimate  : 0.45
```

Figure 2. Skewness and kurtosis results for each cluster

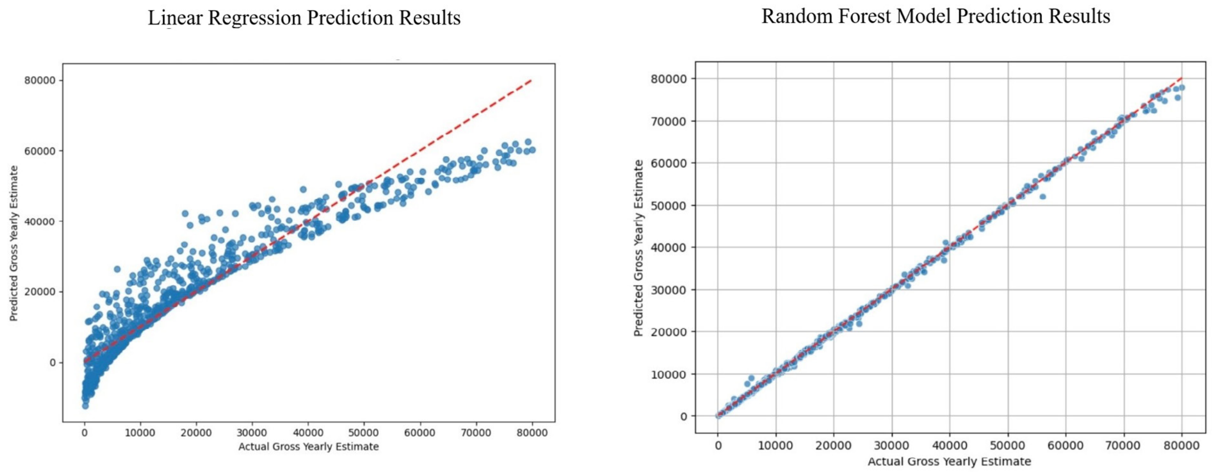

Figure 3. Comparison between linear regression prediction and random forest model prediction results

4.1. Performance evaluation of machine learning models

The performance of the predictive models was evaluated to assess how effectively they estimated yearly revenue for Edinburgh Airbnb listings. As shown in Table 3, the Random Forest Model (RFM) achieved a Pearson correlation coefficient of 0.9988 (p < 0.001) between predicted and actual values, indicating an exceptionally strong linear association and robust model integrity, consistent with the criteria outlined by Montgomery et al. (2012), Hong and Lynn (2020), and Benesty et al. (2009). The high R^2 scores further demonstrate strong predictive accuracy, suggesting that the model reliably captures systematic variation in revenue rather than random noise.

To understand which variables exert the greatest influence on the model's predictions, the feature-importance rankings generated by the RFM were examined. These impurity-based importance scores show that nightly price, availability, and guest capacity consistently emerge as the strongest predictors of gross yearly revenue, with neighbourhood characteristics and review activity contributing to a lesser extent. This provides empirical validation that these attributes are not only intuitively relevant but also quantitatively central to the model's structure.

In addition, to ensure that reductions in prediction error within clusters were not simply a mechanical result of narrowing the revenue range, a revenue-only baseline comparison was conducted. Listings were divided into three terciles of observed gross yearly revenue, and the same Linear Regression and Random Forest models were re-estimated within each subset (Table 4). As expected, the revenue-defined groups produced lower RMSE values (RF: £147–£941) due to their restricted target distribution. In contrast, the K-Means clusters exhibited higher RMSE ranges (RF: £653–£2,617); however, these segments capture multidimensional property structures, such as availability, bedrooms, bathrooms, accommodates, price, and location, that revenue-based bins cannot reveal. This confirms that clustering provides meaningful structural segmentation rather than a discretised form of the outcome variable, reinforcing its role within the overall analytical framework.

Table 3. Correlation analysis and predictive accuracy of RFM model across clusters

Cluster 0	Linear regression: R^2 = 0.98 RSME = 2,599	Random forest model: R^2 = 0.9978 RSME=725.21
Cluster 1	Linear regression: R^2 = 0.96 RSME = 2,882	Random forest model: R^2 = 0.9966 RSME = 798.03
Cluster 2	Linear regression: R^2 = 0.91 RSME = 1587.93	Random forest model: R^2 = 0.9980 RSME = 227.97

Table 4. RMSE comparison across grouping methods

Grouping method	Tier	LR RMSE (£)	RF RMSE (£)
Revenue terciles	Low	1009.43	147.26
	Mid	3371.14	575.24
	High	4898.48	941.5
K-means clusters	C0	3693.63	1292.89
	C1	2035.44	653.68
	C2	5438.54	2617.21
No grouping (full sample)	—	6817.81	679.19

4.2. Overview of clustering results and investment tiers

K-Means clustering was employed to group each Airbnb property in Edinburgh into three distinct clusters: low, medium and high earnings potential for investment. This approach enables the grouping of properties based on shared characteristics, including price, gross yearly estimates, neighbourhood, availability, property size and type.

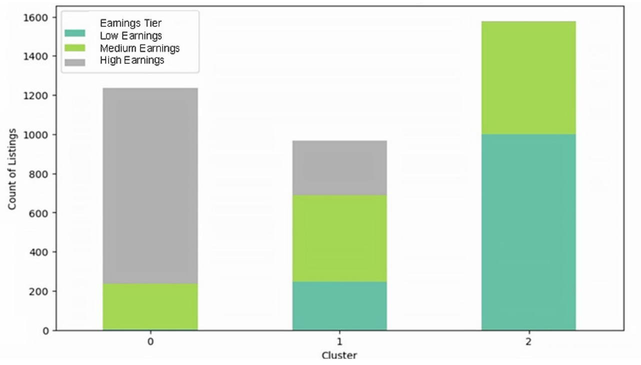

Figure 4. Stacked bar plot of earnings tiers per cluster

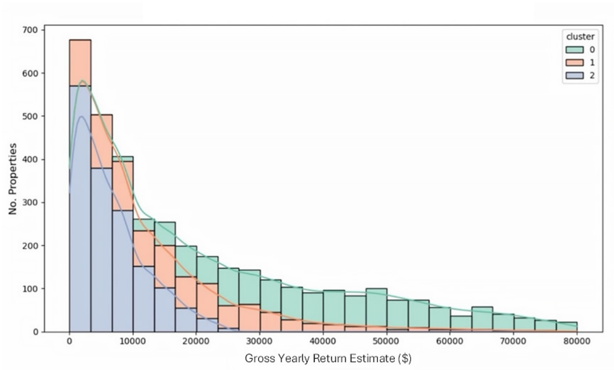

Figure 5. Gross yearly return estimate by cluster

Each cluster captures properties with similar earnings potential, offering valuable insights depending on the investor's objectives. Notably, significant variation exists across clusters in terms of revenue, property type, and investment tier distribution, as shown in Figure 4.

Cluster 0 includes predominantly properties with high earnings potential, i.e. properties that can earn stable and predictable revenue for investors.

Cluster 1 has a balanced mix of properties with medium and high earnings potential, indicating moderate investment risk and moderate return for investors.

Cluster 2 includes predominantly properties with low earnings potential, while investment properties with medium earnings potential form the second largest category. This cluster reflects higher variability and potentially higher investment risk.

Further insights are provided by the histogram of gross yearly estimates. While Clusters 0 and 1 exhibit right-skewed distributions, reflecting higher revenue potential for certain listings, Cluster 2 demonstrates consistently lower revenue, consistent with its low-risk investment profile.

4.3. Revenue trends

Figure 5 showcases the variation in gross yearly estimates across clusters. Cluster 0 demonstrates the highest earnings potential, with revenue ranging from £30,000 to £50,000. Cluster 1 shows lower turnover around £20,000. Due to the diverse property characteristics within the cluster, there is a broader range of revenue estimates. Cluster 2 records the lowest earnings potential, with most properties generating less than £10,000 per year.

4.4. Property type and availability

Figure 6 illustrates the distribution of property types across clusters. Cluster 0 and 1 are dominated by entire homes and apartments, with some private rooms in shared accommodations. The property mix in Cluster 0, consistent with its high-earnings profile, maintains a high average availability of around 217 days per year. Cluster 1, while also offering primarily entire units, shows a markedly lower average availability of 77 days, suggesting a more seasonal offering. In contrast, Cluster 2 shows greater diversity in property types than Cluster 0 and 1, with the largest presence of shared accommodation, e.g. a private room in shared accommodation, albeit with a large presence of entire homes. Cluster 2 shows the lowest availability, offering on average less than 70 days per year.

The data strongly suggests Cluster 0 targets travellers seeking privacy and comfort, while Cluster 1 likely appeals to more budget-conscious visitors, and Cluster 2 accommodates a broad range of needs. These distinctions are illustrated in Figure 7.

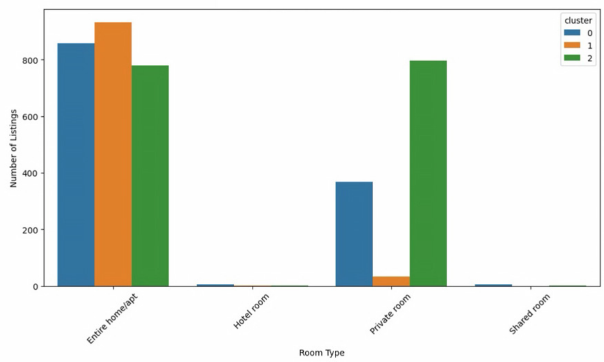

Figure 6. Property type distribution by cluster

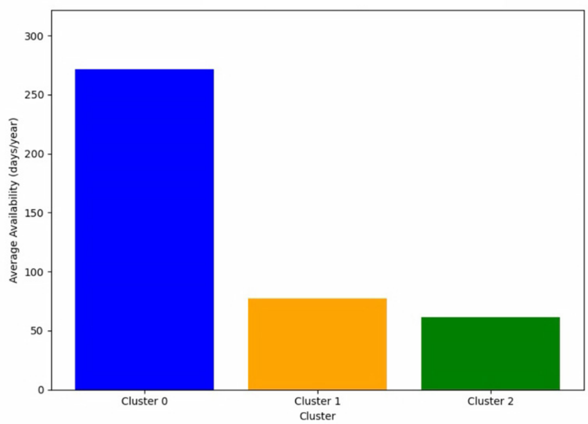

Figure 7. Average availability per cluster

4.5. Price per night

As shown in Figure 8, the analysis demonstrates a variation in the price per night across clusters with averages representing flat rates and not accounting for dynamic pricing strategies. Cluster 0 has the highest average price at £180 per night, which reflects the premium locations and property types. Cluster 1 follows with an average of £150 per night, which aligns with the premium positioning in desirable areas relative to lower-tier clusters. Cluster 2, with an average price of £125, predominantly features shared accommodations, indicating a focus on budget-conscious travellers. It is worth noting that Cluster 2 exhibited the greatest variability in property type, location, and size, with some outliers representing larger group accommodations.

The following cluster-specific analysis provides a detailed breakdown of the neighbourhood locations, property type and predicted revenue for each cluster, as illustrated in Figure 9.

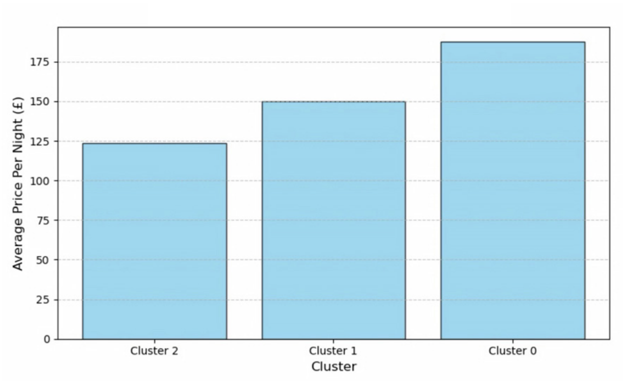

Figure 8. Average price per night by cluster

Cluster 0
High Earnings Potential (Premium Market Category)
Full-time Investment Model

Designed for full-time investors aiming for high revenue, consistent occupancy, and long-term stability. This cluster is ideal for those seeking to operate a professional, year-round Airbnb business.

- Projected gross yearly income: £30,000–£50,000
- Key locations: Old Town, New Town, Princes Street and other high-demand tourist zones
- Property type: Entire homes and apartments
- Average availability: 217+ days per year

Cluster 1

This model is ideal for investors pursuing a balanced risk-and-return strategy, requiring moderate capital outlay compared to Cluster 0. It also appeals to hosts seeking seasonal flexibility, allowing for personal use of the property during off-peak periods.

- Projected average gross yearly income: £20,000
- Locations: Desirable but less central locations, such as Leith and Stockbridge
- Property type: Entire homes and apartments, private rooms
- Average availability: Average of 77 days per year

Cluster 2

This cluster is ideal for Airbnb hosts seeking a low-risk investment with low potential return. Characterised by significant variability in property type and location, it primarily features private rooms in shared accommodations, appealing to budget-conscious travellers. Although shared spaces dominate, there is also a notable presence of entire homes and apartments.

- Projected gross yearly income: £10,000
- Locations: Residential areas and outer districts

Figure 9. Comparison of price per night and property types across clusters

- Property type: The majority are shared accommodations or private rooms
- Recommended availability: Average of 70 days per year

Overall, the analysis shows price variations across clusters. Cluster 0 has the highest average price of £180 per night, reflecting its premium locations and property types. Cluster 1 follows with £150 per night, targeting a balanced risk-reward profile. Cluster 2, with the lowest average price of £125, caters mostly to budget travellers interested in shared accommodations.

Cluster 0, with a projected yearly income of £30,000 to £50,000, is ideal for full-time investors seeking high revenue and steady occupancy. Cluster 1 offers moderate earnings (£20,000) and is suited for seasonal hosting. Cluster 2, with a projected income of £10,000, is best for entry-level investors, offering low-risk opportunities in shared accommodation with 70 days of availability per year.

5. Discussion

The analysis of Edinburgh's Airbnb market reveals important insights into the dynamics of investment opportunities in properties, potential revenue generation and property characteristics across distinct segments. Through the application of machine learning models and statistical techniques, including K-Means clustering, linear regression, and Random Forest modelling, a nuanced understanding of the market was developed.

5.1. Model performance and predictive capability

Initial attempts to model revenue using a single linear regression model yielded modest predictive accuracy, with an R^2 of 0.72 and an RMSE of £5,800. While this R^2 value is statistically reasonable, it suggests the model could only account for 72% of the variance in gross yearly earnings, which is insufficient for robust forecasting, especially given the financial implications for investors. Furthermore, the RMSE indicated considerable absolute prediction error, implying significant variability in property-level revenue that the linear model struggled to capture. This is consistent with findings by Montgomery et al. (2012) who emphasise the limitations of linear models when applied to heterogeneous datasets. Additionally, skewness and kurtosis values falling outside expected ranges in some clusters further indicate departures from normality, complicating the assumptions underlying linear models.

In contrast, the Random Forest Model demonstrated a significant improvement in predictive accuracy. When applied to the full dataset, the model achieved an R^2 of 0.9994 and a substantially lower RMSE of £470. Cluster-level analysis reinforced this finding, with the Random Forest consistently outperforming linear regression across all clusters, particularly excelling in Cluster 2 where linear regression had struggled. These results are consistent with Hong and Lynn (2020) work highlighting the effectiveness of ensemble methods in capturing non-linear patterns in property data.

The Random Forest Model also identified availability, price and guest capacity as the most important predictors of gross yearly revenue. This aligns with prior research, suggesting that consistent availability, competitive pricing and the ability to accommodate more guests are key drivers of Airbnb profitability.

Even though the data shows that each cluster has properties residing across all locations in Edinburgh, there are distinct patterns in the data regarding ownership structures, property availability and financial predictability, which is how the clusters were formed.

The term risk in this project refers to the predictability of revenue estimations based on price per night, location, occupancy limits and property type. For example, low-risk properties in desired areas with predictable revenue offer greater financial security.

Cluster 0 is considered high earnings potential and a low-risk investment compared to clusters 1 and 2 because the data showed stable revenue across the properties that are predominantly located in high-demand areas, such as Edinburgh's Old Town.

The data also shows properties in Cluster 0 are almost exclusively entire homes or apartments with high year-round availability, suggesting these are properties run solely for business use and not as primary residences. The application of the RFM showed that properties in Cluster 0 exhibit minimal revenue fluctuation and consistent occupancy rates. Combined with historical data showing higher prices per night, optimal availability and location, the description of properties in Cluster 0 fits the low-risk investment category. This also suggests that any Airbnb listing that aligns with the Cluster 0 characteristics is likely to achieve predictable earnings. Potential Airbnb investors looking at this type of investment will face higher scrutiny and licencing costs, given most are located in Edinburgh's "Control Area" zones and are secondary letting properties.

Unlike Cluster 0, Cluster 1 offers a more diverse range of properties in terms of ownership and rental strategies. While there are strong similarities between clusters 0 and 1 in terms of property desirability and type, the lower availability offered in Cluster 1 suggests these are "part-time" rentals. Most likely, these Airbnb listings are operated by hosts who rent out their properties during peak seasons, such as the Christmas Markets or the summer Fringe Festival. As a result, Cluster 1 presents a moderate investment risk where sustainable revenue can still be achieved if hosts maximise earnings during peak periods.

The data for Cluster 2 shows most variability and unpredictable returns. Cluster 2 is the most diverse of all three clusters in terms of property types, location, price per night and rental patterns, which makes it the least predictable in terms of revenue generation. A notable distinction of Cluster 2 is the high popularity of shared accommodation, compared to other clusters. This suggests the guest demographic is looking for budget accommodation, instead of a private apartment. However, private homes and apartments are the second largest property type, with outliers in the data showing properties with a large number of bedrooms and bathrooms. This means that these properties in Cluster 2 cater to larger parties. Cluster 2 is considered a high-risk, low-earnings opportunity due to the revenue unpredictability.

5.2. Government regulation of short-term rentals

In 2021, the Scottish Government implemented new regulations regarding short-term rentals. This was in response to the growing impact short-term rentals have on housing availability and affordability. New regulations include the Business and Regulatory Impact Assessment (BRIA) relating to the Civic Government (Scotland) Act 1982 (Licensing of Short-term Lets) Order 2021 ("the Licensing Order") and the Town and Country Planning (Short-Term Let Control Areas) (Scotland) Regulations 2021 ("the Control Area Regulations"). In a document released by Airbnb in 2020 detailing the steps Airbnb hosts must take to comply with these regulations, Airbnb (2020) described Scotland's short-term lets regulation proposal as "costly, complex and unfair".

After noticing a slightly negative impact on the housing market and local economy, the Scottish Government acted to avoid the disruption Airbnb has caused in other European cities, such as those discussed by Xu et al. (2019), Cocola-Gant and Gago (2021). The City of Edinburgh Council's (2023) most recent report argues that the property market is stable. Table 5 summarises the licencing requirements for various properties and letting options.

The regulatory framework for short-term rentals in Scotland is stringent, with full compliance mandated by law.

6. Conclusions

This study has leveraged K-Means Clustering and the application of Random Forest Model and Linear Regression Model to successfully categorise the data into investment tiers, each predicting revenue, affordability and showing implications of regulation. The RFM results quantified revenue potential, which ensured that data guided the investment tiers based on potential earnings. This allows prospective investors to make data-based decisions regarding operating a property offered for rent via Airbnb in the Edinburgh short-term rental market.

The clustering results showed three distinct investment tiers:

- Cluster 0 is considered an attractive investment opportunity due to its high earnings potential. The data showed stable and predictable revenue of £30,000 to £50,000 per year due to being a "full-time" Airbnb rental property, offering guests full homes and apartments in the most desirable areas. Short-term rentals that match the categorisation of Cluster 0 must abide by strict regulations and are subject to costly licencing fees.
- Cluster 1 is considered a balanced investment with a modest earnings potential of £20,000 per year. However, much like Cluster 0 properties in terms of type and location, Cluster 1 offers less availability for properties renting out via Airbnb on a part-time basis, which offers seasonal availability, during peak times. Short-term rentals that match the categorisation of Cluster 1 must also abide by strict regulations, and the high fees may not be cost-effective for the limited availability offered throughout the year.
- Cluster 2 is considered an investment at the lower tier with unpredictable revenue potential due to the diverse range of property types, which suggests a diverse range of travellers. The categorisation of the Cluster 2 results suggests that the target audience is a budget-oriented guest demographic due to the popularity of shared accommodation in suburban areas. The other part of Cluster 2 caters to a different

Table 5. Licencing requirements for accommodations in Edinburgh

Licence required before operating	▪ New hosts must secure a licence before accepting bookings for their Airbnb ▪ Non-compliance is a criminal offence			
Licence types	Home sharing Renting out rooms in a primary residence	Home letting Renting out a primary home while absent	Secondary letting Renting out a property that is *not* a primary residence	Combination of Home letting & Homes sharing
Mandatory safety & compliance measures	Fire, gas & electrical safety checks	Legionella risk assessments	Carbon monoxide & smoke detectors	Proof of insurance
Additional conditions set by local authorities	▪ Edinburgh City is a 'Control area' that requires additional planning permission for secondary lets ▪ Limits on maximum occupancy per accommodation ▪ Restrictions on property modifications (e.g. adding additional beds) ▪ Possible noise monitoring requirements			
Licence fees & renewal	▪ Licences are valid for up to 3 years and must be renewed ▪ Fees can vary based on property size, location and number of guests			

target audience: entire homes and apartments with many bedrooms and bathrooms are attractive for tourists staying in Edinburgh in large groups. Regulatory implications for short-term rentals in this category depend on the variations mentioned.

When applied to each cluster, the Random Forest Model outperformed the Linear Regression Model by achieving higher predictive accuracy. Cluster 0 achieved R^2 of 0.9978 and an RSME of £725.21; Cluster 1 achieved R^2 of 0.9966 and an RSME of £798.03; Cluster 2 achieved R^2 of 0.9978 and an RSME of £725.21.

These results help to confirm that potential revenue across the Edinburgh Airbnb listings is not random and can be accurately predicted based on key listing characteristics. The paper contributes to knowledge by developing a framework that supports potential investors to maximise their return on investment in an increasingly competitive short-term rental market and shared tourism accommodation economy. When applied to the K-Means clustering results, the Random Forest Model showed that pricing, availability and guest capacity were the variables that influence revenue most. Incorporating predictive modelling into the market categorisation enabled this study to offer a practical, data-based framework for potential investors who are considering various options for investment in properties with an intent to offer them for short-term rent via Airbnb in Edinburgh.

Limitations of the study

The study relies on publicly available Inside Airbnb data representing the Edinburgh market as of September 2023. As a single-year snapshot, more recent market changes are not captured, which limits the findings' long-term applicability.

The analysis focuses on nightly price as the main revenue metric. Broader market forces, such as property values, wider economic trends or regulatory shifts, are not included and fall outside the study's scope. The findings therefore apply primarily to Edinburgh and may not generalise to other locations without adjustment.

The dataset does not include daily booking calendars, monthly occupancy patterns or dynamic pricing histories. As a result, the study assumes constant nightly price and average occupancy. Seasonal fluctuations and event-driven price changes cannot be modelled, and revenue values should be interpreted as baseline estimates.

Because the dataset reflects conditions during the post-pandemic recovery period, revenue and occupancy patterns from 2023 may not represent long-term market behaviour. Results should be understood within this specific temporal context.

Location is represented using neighbourhood codes, which offer basic spatial differentiation but do not capture spatial autocorrelation. Methods such as geographically weighted regression or spatial clustering were not feasible due to the lack of coordinate-level data.

Suggestions for further research

Given the complexity of the short-term rental market, future studies could focus on a dynamic pricing strategy and sentiment analysis, among many other options.

Dynamic pricing accounts for occupancy trends and adjusts pricing according to booking rates, competitor bookings and seasonality. The research of Abrate et al. (2022), Gibbs et al. (2018) and Leoni and Nilsson (2021) explores different approaches to dynamic pricing for Airbnb business for both professional and individual hosts. The studies found a positive correlation between the use of dynamic pricing models and optimised revenue for hosts engaging in dynamic pricing strategies, drastically increasing their potential revenue. Implementing dynamic pricing in conjunction with the cluster tiers and predicted revenue may ensure an enhanced revenue outcome, which needs to be investigated.

Sentiment analysis would add depth to future studies by drawing insights into the individual Airbnb listings in each cluster and also by incorporating the geographical importance of the micro-environment in Edinburgh. Xu et al. (2019) study on the geographical micro-environment of Airbnb supports this suggestion. A closer look at the consumer behaviour regarding Airbnb accommodations and case studies involving current Airbnb hosts in popular tourist cities would be useful for a more nuanced assessment of investment opportunities in properties offered for short-term rent.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Author contributions

NN: Involved in the initial conception and design of the study; contributed to data acquisition, performed data analysis, and assisted in interpreting findings. EO: Participated in developing the study design and methodology; contributed to data analysis; helped shape the interpretation of results; wrote sections of the manuscript and reviewed it; approved the final version. MN: Assisted in study conception and planning; contributed to study design and data analysis; played a role in synthesizing and interpreting results; prepared portions of the manuscript and engaged in substantive editing; approved the final version.

Disclosure statement

The authors declare that they have no competing financial, professional, or personal interests that could have influenced the work reported in this article.

References

Abrate, G., Sainaghi, R., & Mauri, A. G. (2022). Dynamic pricing in Airbnb: Individual versus professional hosts. *Journal of Business Research*, *141*, 191–199. https://doi.org/10.1016/j.jbusres.2021.12.012

Airbnb. (2020). *Scotland short-term lets regulation proposals*: *The steps you'll need to take to comply*. https://news.airbnb.com/wp-content/uploads/sites/4/2020/10/ScotlandShortTermLetsProposal1310.pdf

Airbnb. (2024). *Six months on*: *An update on the impact of Scotland's short-term let rules*. https://news.airbnb.com/en-uk/six-months-on-an-update-on-the-impact-of-scotlands-short-term-let-rules/

Barron, K., Kung, E., & Proserpio, D. (2020). The effect of home-sharing on house prices and rents: Evidence from Airbnb. *Marketing Science*, *40*(1), 23–47. https://doi.org/10.1287/mksc.2020.1227

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (Eds.). (2009). Pearson correlation coefficient. In *Noise reduction in speech processing*: *Vol. 2*. *Springer topics in signal processing* (pp. 1–4). Springer. https://doi.org/10.1007/978-3-642-00296-0_5

Bibler, A., Teltser, K., & Tremblay, M. (2021). *Is sharing really caring? The effect of Airbnb on the affordability of housing*. SSRN. https://doi.org/10.2139/ssrn.3846919

Camatti, N., di Tollo, G., Filograsso, G., & Ghilardi, S. (2024). Predicting Airbnb pricing: A comparative analysis of artificial intelligence and traditional approaches. *Computational Management Science*, *21*, Article 30. https://doi.org/10.1007/s10287-024-00511-4

Cervera, D. D. J., de Esteban Curiel, J., & Pérez-Bustamante Yábar, D. C. (2024). Machine learning for short-term property rental pricing based on seasonality and proximity to food establishments. *British Food Journal*, *126*(13), 332–352. https://doi.org/10.1108/BFJ-07-2023-0634

Chen, L.-P. (2021). Practical statistics for data scientists: 50+ essential concepts using r and python. *Technometrics*, *63*(2), 272–273. https://doi.org/10.1080/00401706.2021.1904738

Cocola-Gant, A., & Gago, A. (2021). Airbnb, buy-to-let investment and tourism-driven displacement: A case study in Lisbon. *Environment and Planning A*: *Economy and Space*, *53*(7), 1671–1688. https://doi.org/10.1177/0308518X19869012

Congiu, R., Pino, F., & Rondi, L. (2024). The uneven effect of Airbnb on the housing market: Evidence across and within Italian cities. *Journal of Regional Science*, *65*(2), 339–377. https://doi.org/10.1111/jors.12737

Ding, K., Niu, Y., & Choo, W. C. (2023). The evolution of Airbnb research: A systematic literature review using structural topic modelling. *Heliyon*, *9*(6), Article e17090. https://doi.org/10.1016/j.heliyon.2023.e17090

Duso, T., Michelsen, C., Schäfer, M., & Tran, K. D. (2024). Airbnb and rental markets: Evidence from Berlin. *Regional Science and Urban Economics*, *106*, Article 104007. https://doi.org/10.1016/j.regsciurbeco.2024.104007

Edinburgh Council. (2023). *Statement of licensing policy*. https://consultationhub.edinburgh.gov.uk/ce/overprovision-of-licensed-premises/user_uploads/statement_of_licensing_policy_2023-1.pdf

Gabriel, F. S., Ribeiro, K. D. S., & Rogers, P. (2015). Real estate investment trusts performance: Brazil versus United States. *Business and Management Review*, *1*, 45–67.

García-López, M. À., Jofre-Monseny, J., Martínez-Mazza, R., & Segú, M. (2020). Do short-term rental platforms affect housing markets? Evidence from Airbnb in Barcelona. *Journal of Urban Economics*, *119*, Article 103278. https://doi.org/10.1016/j.jue.2020.103278

Gauß, P., Gensler, S., Kortenhaus, M., Riedel, N., & Schneider, A. (2024). Regulating the sharing economy: The effects of day caps on short- and long-term rental markets and stakeholder outcomes. *Journal of the Academy of Marketing Science*, *52*, 1627–1650. https://doi.org/10.1007/s11747-024-01028-7

Gibbs, C., Guttentag, D., Gretzel, U., Yao, L., & Morton, J. (2018). Use of dynamic pricing strategies by Airbnb hosts. *International Journal of Contemporary Hospitality Management*, *30*(1), 2–20. https://doi.org/10.1108/IJCHM-09-2016-0540

Guttentag, D. A., Mahdikhani, M., Starr, C., & Starr, E. (2026). How Airbnb guest experiences differ by accommodation type: A lexicon-based sentiment analysis of Airbnb reviews. *Tourism Recreation Research*, *51*(1), 1–15. https://doi.org/10.1080/02508281.2025.2449627

Hong, S., & Lynn, H. S. (2020). Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. *BMC Med Research Methodology*, *20*, Article 199. https://doi.org/10.1186/s12874-020-01080-1

Hur, D., Lee, S., & Kim, H. (2024). The impact of Airbnb on long-term rental markets in San Francisco: A geospatial analysis using multiscale geographically weighted regression. *International Journal of Geo-Information*, *13*(9), Article 298. https://doi.org/10.3390/ijgi13090298

Iliopoulou, P., Krassanakis, V., & Kappelos, K. (2025). Modeling the spatial impact of short-term rentals on house prices: The case of Athens, Greece. *Urban Science*, *9*(12), Article 539. https://doi.org/10.3390/urbansci9120539

Inside Airbnb. (2024). *Get the data*. https://insideairbnb.com/get-the-data/

Jaroszewicz, J., & Horynek, H. (2024). Aggregated housing price predictions with no information about structural attributes – hedonic models: Linear regression and a machine learning approach. *Land*, *13*(11), Article 1881. https://doi.org/10.3390/land13111881

Jin, G. Z., Wagman, L., & Zhong, M. (2024). The effects of short-term rental regulation: Insights from Chicago. *International Journal of Industrial Organization*, *96*, Article 103087. https://doi.org/10.1016/j.ijindorg.2024.103087

Koster, H. R. A., van Ommeren, J., & Volkhausen, N. (2021). Short-term rentals and the housing market: Quasi-experimental evidence from Airbnb in Los Angeles. *Journal of Urban Economics*, *124*, Article 103356. https://doi.org/10.1016/j.jue.2021.103356

Lee, H., Han, H., Pettit, C., Gao, Q., & Shi, V. (2024). Machine learning approach to residential valuation: A convolutional neural network model for geographic variation. *Annals of Regional Science*, *72*, 579–599. https://doi.org/10.1007/s00168-023-01212-7

Lee, S., & Kim, H. (2023). Four shades of Airbnb and its impact on locals: A spatiotemporal analysis of Airbnb, rent, housing prices, and gentrification. *Tourism Management Perspectives*, *49*, Article 101192. https://doi.org/10.1016/j.tmp.2023.101192

Leoni, V., & Nilsson, W. (2021). Dynamic pricing and revenues of Airbnb listings: Estimating heterogeneous causal effects. *International Journal of Hospitality Management*, *95*, Article 102914. https://doi.org/10.1016/j.ijhm.2021.102914

Li, H., & Srinivasan, K. (2019). Competitive dynamics in the sharing economy: An analysis with a focus on Airbnb and hotels. *Marketing Science*, *38*(3), 365–391. https://doi.org/10.1287/mksc.2018.1143

Liu, L., Yu, H., Zhao, J., Wu, H., Peng, Z., & Wang, R. (2022). Multiscale effects of multimodal public facilities accessibility on housing prices based on MGWR: A case study of Wuhan, China. *ISPRS International Journal of Geo-Information*, *11*(1), Article 57. https://doi.org/10.3390/ijgi11010057

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, *23*(1), 151–169. https://doi.org/10.1146/annurev.publhealth.23.100901.140546

Mathotaarachchi, K. V., Hasan, R., & Mahmood, S. (2024). Advanced machine learning techniques for predictive modeling of property prices. *Information*, *15*(6), Article 295. https://doi.org/10.3390/info15060295

Montgomery, D. C., Peck, E. A., & Vining, G.G. (2012). *Introduction to linear regression analysis* (5th ed.). Wiley.

Nieuwland, S., & van Melik, R. (2020). Regulating Airbnb: How cities deal with perceived negative externalities of short-term rentals. *Current Issues in Tourism*, *23*(7), 811–825. https://doi.org/10.1080/13683500.2018.1504899

Oskam, J. A. (2019). *The future of Airbnb and the 'sharing economy': The collaborative consumption of our cities*. Channel View Publications. https://doi.org/10.21832/9781845416744

Rodriguez-Serrano, J. A. (2025). Prototype-based learning for real estate valuation: A machine learning model that explains prices. *Annals of Operations Research*, *344*, 287–311. https://doi.org/10.1007/s10479-024-06273-1

Rossi, F., & d'Addona, S. (2025). The impact of Airbnb on long-term rentals, population, and college enrollments: Empirical evidence from an Italian university town. *Tourism Economics*, *31*(8), 1674–1701. https://doi.org/10.1177/13548166251314063

Serrano, L., Sianes, A., & Ariza-Montes, A. (2020). Understanding the implementation of Airbnb in urban contexts: Towards a categorization of European cities. *Land*, *9*(12), Article 522. https://doi.org/10.3390/land9120522

Sharma, H., Harsora, H., & Ogunleye, B. (2024). An optimal house price prediction algorithm: XGBoost. *Analytics*, *3*(1), 30–45. https://doi.org/10.3390/analytics3010003

Tang, J., Cheng, J., & Zhang, M. (2024). Forecasting Airbnb prices through machine learning. *Managerial and Decision Economics*, *45*(1), 148–160. https://doi.org/10.1002/mde.3985

Wang, Z., Wang, J., Wu, S., & Du, Z. (2022). House price valuation model based on geographically neural network weighted regression: The case study of Shenzhen, China. *ISPRS International Journal of Geo-Information*, *11*(8), Article 450. https://doi.org/10.3390/ijgi11080450

Xu, F., Hu, M., La, L., Wang, J., & Huang, C. (2019). The influence of the neighbourhood environment on Airbnb: A geographically weighted regression analysis. *Tourism Geographies*, *22*(1), 192–209. https://doi.org/10.1080/14616688.2019.1586987