Airbnb Price Prediction and Segmentation in New York City

Introduction

Airbnb has transformed the accommodation market since 2008, offering flexible stays that range from spare rooms to entire apartments. Yet in New York City, nightly rates vary widely depending on neighbourhood, property type, and host behaviour. Understanding what drives these differences is essential if Airbnb is to support hosts in setting fair prices while ensuring guests perceive value (Gibbs et al., 2018).

Accurate price prediction could benefit hosts by improving competitiveness and reducing mispricing, and helping Airbnb to forecast demand and identify growth opportunities. While clustering could reveal natural market segments to guide pricing and marketing strategies.

This study uses the Airbnb New York City 2019 dataset from Kaggle (Inside Airbnb, 2019), which contains more than 48,000 listings with information on host activity, location, availability, and price. It asks: *which features most strongly influence nightly price, and how can predictive modelling and clustering be combined to define actionable market segments for Airbnb's strategic planning?*

Modelling Approach

As a baseline, a linear regression model was fitted, providing a transparent way to estimate how predictors relate to nightly price. Coefficients showed the direction and strength of each feature's influence, and diagnostic checks confirmed that assumptions were reasonably satisfied.

To capture non-linear relationships and complex interactions, Random Forest regression was applied to the dataset. This ensemble of decision trees improves predictive accuracy and generates feature importance scores that highlight the strongest drivers of price (Liu et al., 2021; Camatti et al., 2024; Ganta & Krishna, 2024).

Finally, k-means clustering grouped listings by shared attributes such as room type, neighbourhood, reviews, and availability. Clustering revealed natural market segments to inform pricing, marketing, and host strategies (Kadri, 2024). Prior studies support this mix of methods, with reviews confirming that pricing and segmentation are central themes in Airbnb research (Ding et al., 2023; Dolnicar, 2021).

Analysis

EDA

Exploratory data analysis (EDA) revealed skewed distributions. Most listings were priced below $200 per night; however, there were extreme outliers that exceeded $1,000. Minimum night requirements included unrealistic values stretching into years. These were removed to

ensure data quality. Missing values in reviews_per_month were imputed as zero. Variation across categories was clear, with Manhattan and Brooklyn dominating supply and entire apartments commanding higher nightly rates than private or shared rooms.

Linear Regression

Linear regression provided a transparent baseline. After preprocessing with one-hot encoding and scaling, the model explained about 33% of the variance in nightly prices ($R^2 \approx 0.33$) with a mean absolute error of $48.6. Using a log-transformed target improved explanatory power to 52% ($R^2 \approx 0.515$), though back-transformation increased errors due to extreme prices.

The coefficients confirmed that room type was the strongest driver (Figure 1, private rooms were ~55% cheaper, and shared rooms over 70% cheaper than entire apartments. Neighbourhood effects were also pronounced. Manhattan listings achieved an 82% premium relative to the Bronx, followed by Brooklyn and Queens, while Staten Island was associated with lower prices. These findings align with earlier studies highlighting property type and location as key price determinants (Gibbs et al., 2018). Host and availability variables had smaller effects, with multi-listing hosts tending to charge slightly less, and year-round availability being linked to higher prices.

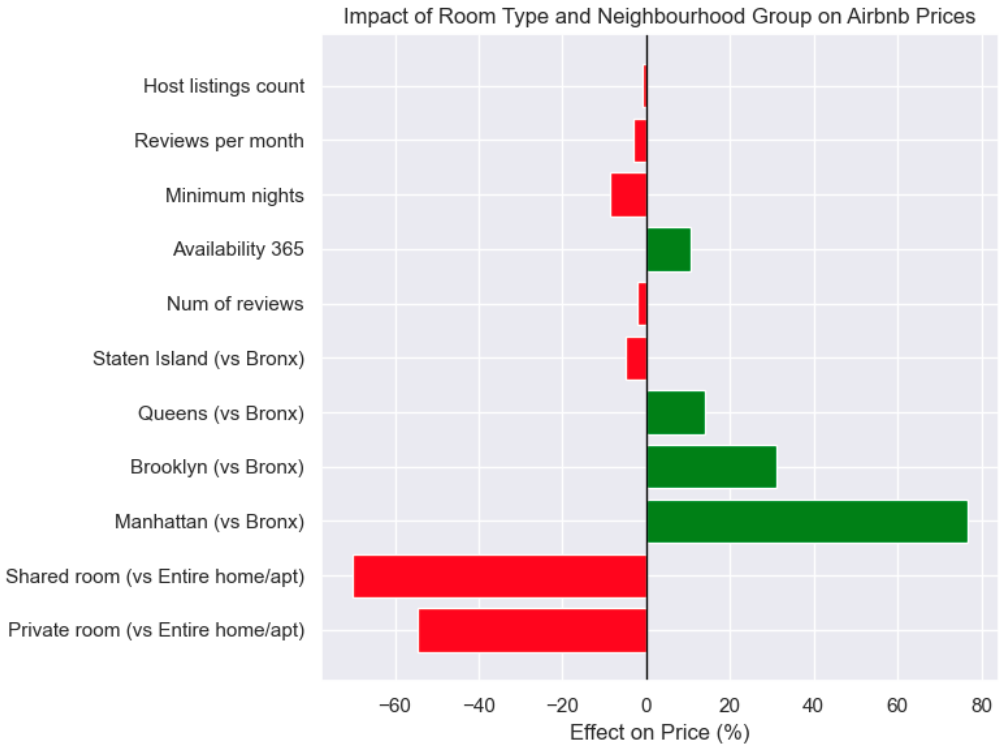

Figure 1

Random Forest

Random Forest regression was then applied to strengthen our analysis by testing whether a non-linear approach would change the results. Here, feature engineering added new explanatory variables such as long-stay indicators, multi-listing host flags, review density, and neighbourhood-level price baselines. A log-transformed target produced the best performance, achieving an R^2 of 0.43 with a mean absolute error of $46.5. Although the predictive gains over linear regression were modest, the model provided valuable insights into feature importance. Location (median neighbourhood price), room type, and availability were the most influential drivers (Figure 2), while review density and host characteristics contributed to a lesser degree. The scatterplot of actual versus predicted prices (Figure 3) showed that the model tracked overall trends but struggled to capture high-price outliers. More advanced ML methods, such as boosting and deep learning, could capture additional non-linearities and improve predictive accuracy (Tang et al., 2024).

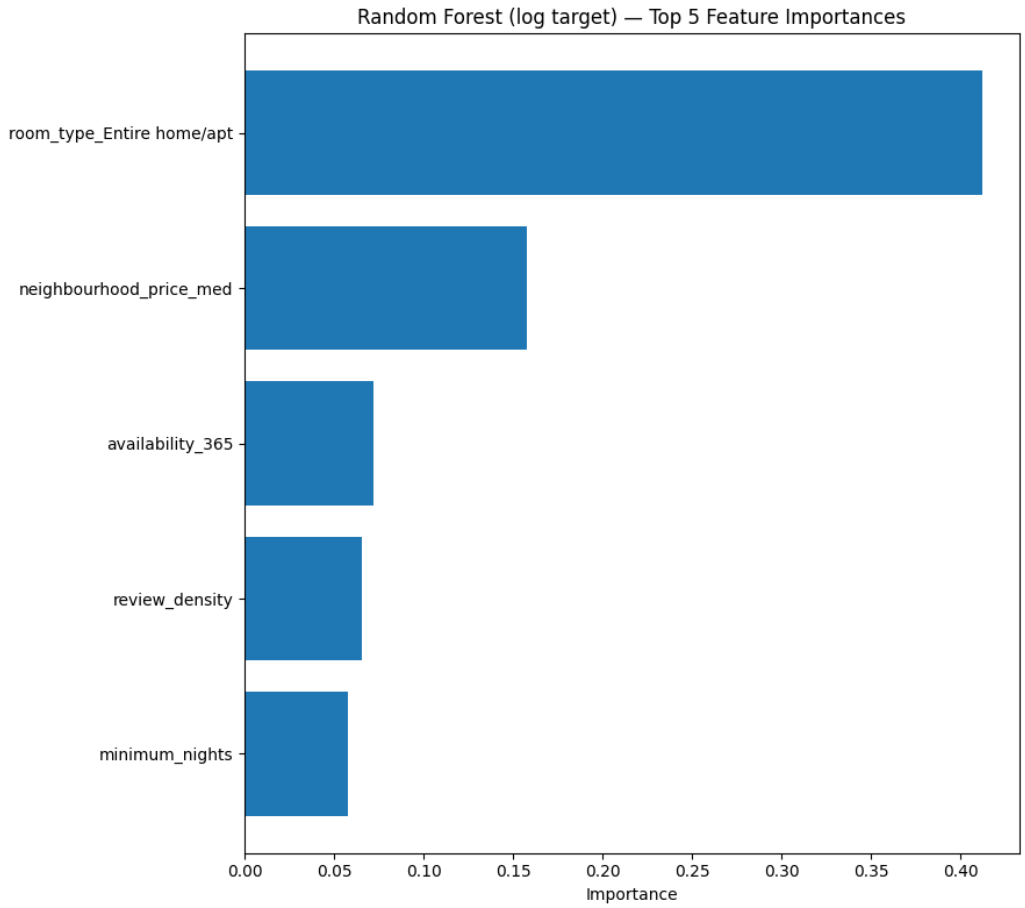

Figure 2

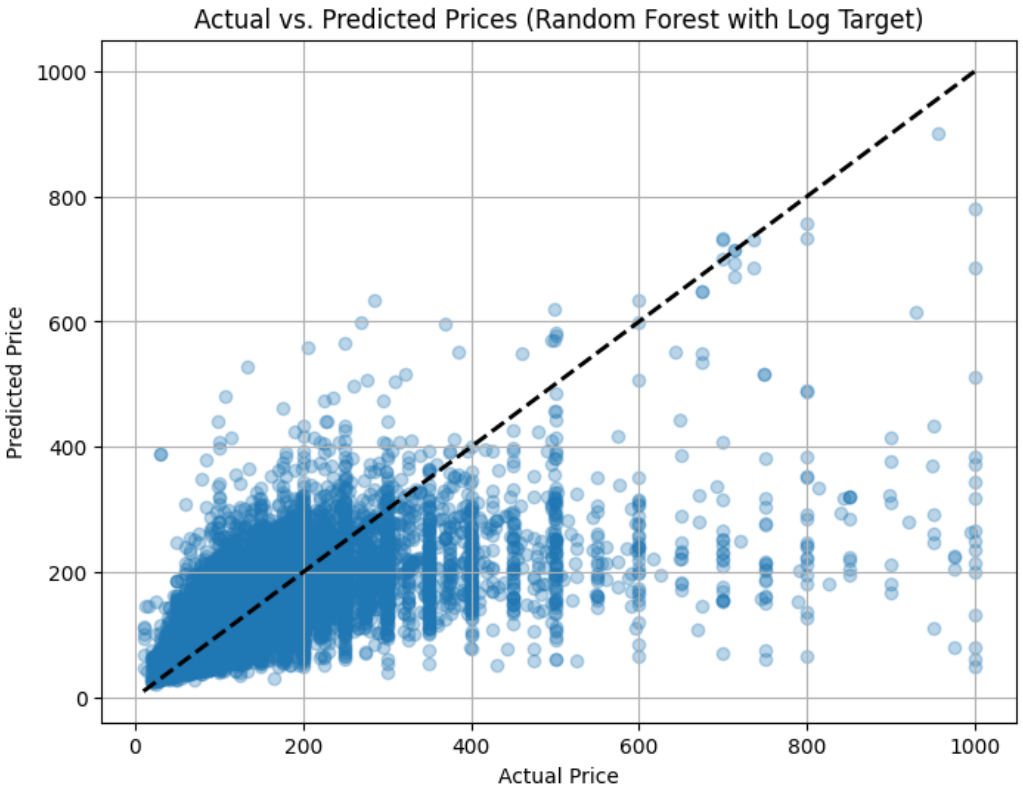

Figure 3

The performance of both models is summarised in Table 1. Linear regression offers interpretability. Random Forest, especially when enhanced with engineered features and a log target, improves accuracy at the expense of transparency.

Model	MAE ($)	RMSE ($)	R^2
Linear Regression (raw price)	48.6	–	0.325
Linear Regression (log price)	~0.35 log points	–	0.515 (log space)
Random Forest (raw price)	55.2	93.2	0.336
Random Forest (log price)	51.0	91.6	0.359
Random Forest (log + features	46.5	86.3	0.431

Table 1. Regression model performance on Airbnb NYC 2019 dataset

Clustering

Clustering was then applied to identify broader market structure. Based on neighbourhood group, room type, and availability, a three-cluster solution emerged (Table 2). The clusters, broadly separated into high-price, high-availability entire homes/apartments (concentrated in Manhattan), moderate mid-market listings (mixed boroughs, varied availability), and budget-oriented private/shared rooms (often in less central boroughs). The scatter and box plots illustrate these distinctions (Figures 4, 5, and 6). For example, Cluster 1 has higher median

prices, especially for entire home/apartment listings in Manhattan. As these features were used in clustering, the plots illustrate their role in defining groups rather than providing independent validation. Still, they nonetheless help give a clear interpretation of the resulting market segments. This complements the regression findings, which identified room type and borough as the primary price levers (Gibbs et al., 2018), and aligns with established approaches to clustering and market segmentation (Jain, 2010; Xu & Wunsch, 2008).

Cluster	Key Characteristics
Cluster 1	High-availability entire homes/apartments; Higher average prices; Concentrated in Manhattan
Cluster 2	Lower-availability listings; Generally lower price points; Mixed borough distribution
Cluster 3	Budget-oriented shared/private rooms; Lower prices; Concentrated in less central boroughs

Table 2. Clustering results: Airbnb NYC 2019 dataset (k=3)

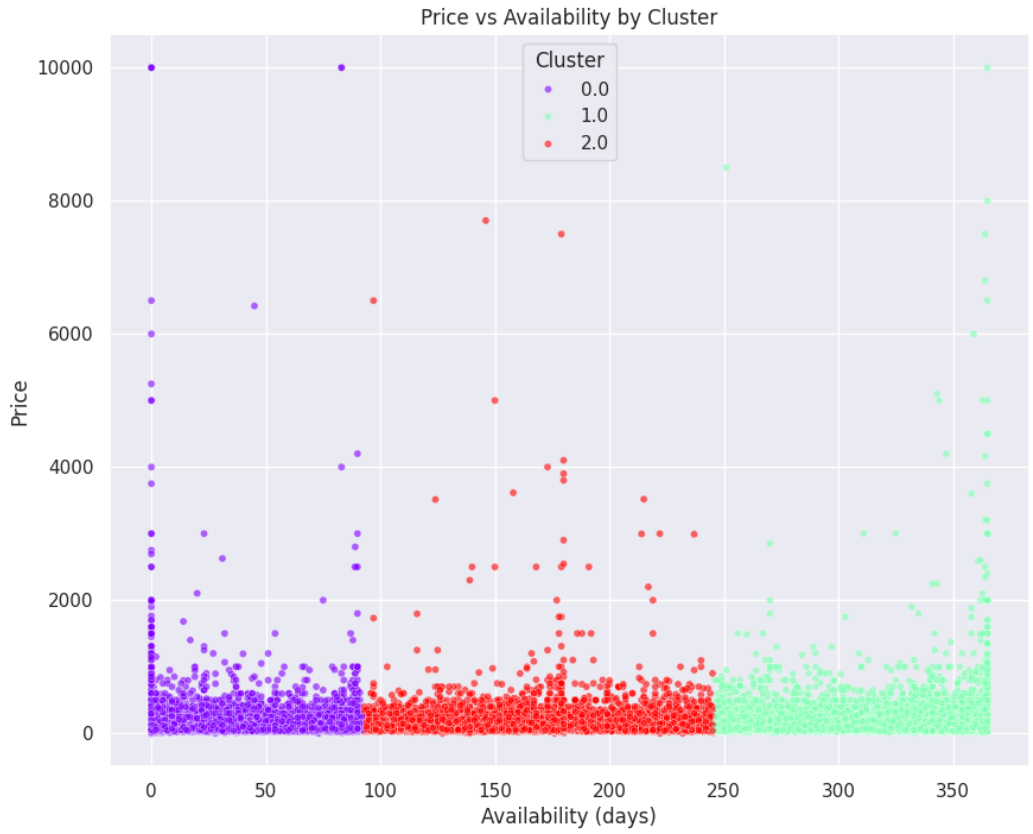

Figure 4

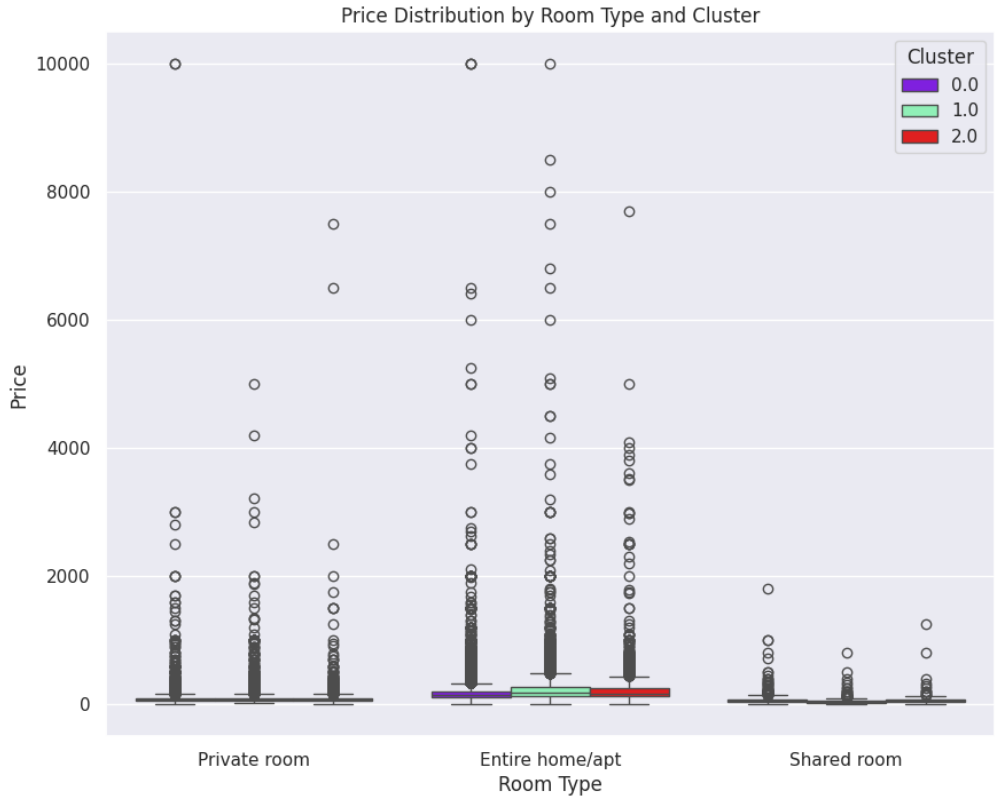

Figure 5

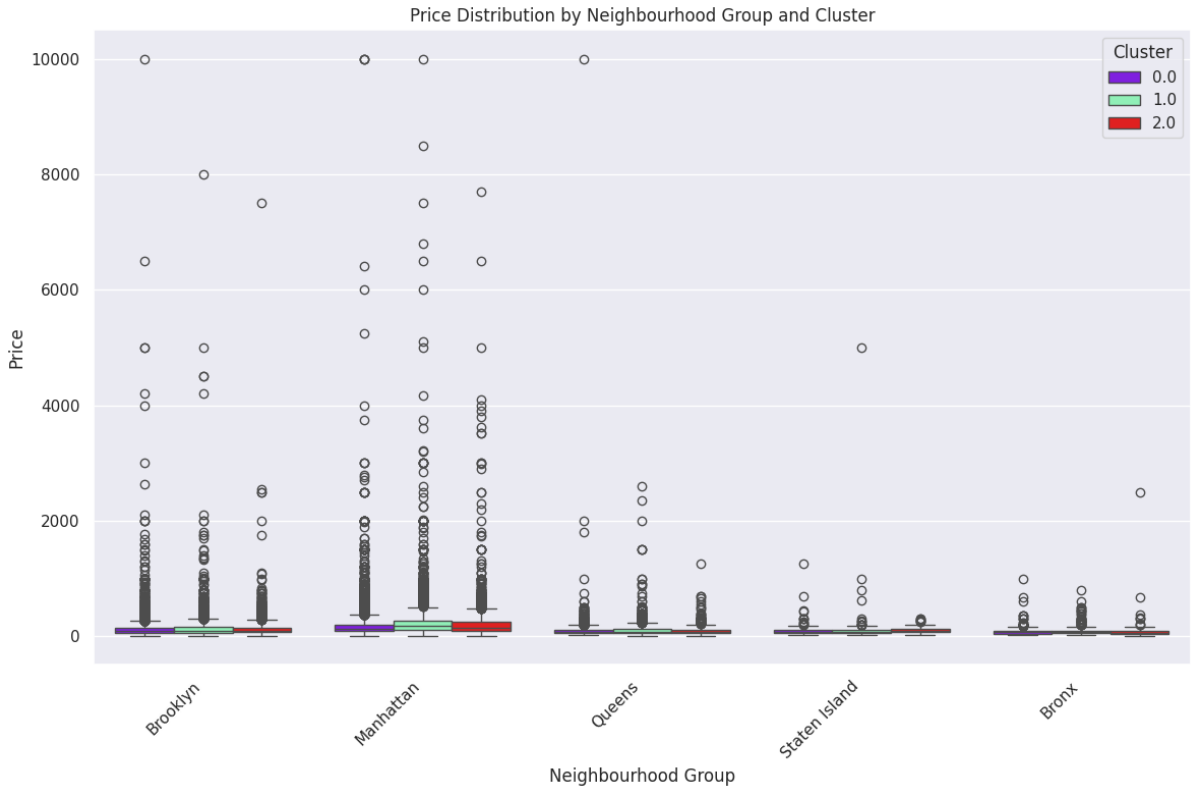

Figure 6

Conclusion

Together, regression quantified the drivers of price and provided predictive benchmarks, while Random Forest highlighted non-linear effects and the value of engineered features. Clustering complemented these results by revealing three actionable market segments for tailored pricing and marketing.

Discussion and Recommendations

This study aimed to determine which features most strongly influence Airbnb prices in New York City and how predictive modelling and clustering can be used together to support strategic planning. The results show that room type and neighbourhood are the dominant drivers of nightly price, with host and availability factors adding secondary effects. Predictive modelling contributes in two ways: regression provides transparency in identifying the key determinants of price, while Random Forest improves accuracy by accounting for non-linear relationships.

Clustering offered a complementary perspective by grouping listings into three distinct market segments defined by availability, location, and pricing level. This segmentation should enable Airbnb to move beyond uniform pricing models, instead tailoring strategies for premium, mid-market, and budget listings (Jain, 2010). Regression and clustering could provide Airbnb with a robust framework for forecasting prices, guiding hosts, and identifying market opportunities. As Dolnicar (2021) highlights, segmentation only delivers value when it is applied to practical decisions such as pricing and communication, underlining the importance of ensuring these insights are used to shape Airbnb's strategic actions.

References

Camatti, N., di Tollo, G., Filograsso, G. and Ghilardi, S., 2024. Predicting Airbnb pricing: a comparative analysis of artificial intelligence and traditional approaches. *Computational Management Science*, 21(1), pp.1–25.

Ding, K., Niu, Y. and Choo, W.C., 2023. The evolution of Airbnb research: a systematic literature review using structural topic modeling. *Heliyon*, 9(6), e17090.

Dogru, T., Hanks, L., Mody, M., Suess, C. and Sirakaya-Turk, E., 2020. *The effects of Airbnb on hotel performance: Evidence from cities beyond the United States. Tourism Management*, 79, 104090. Available at: https://doi.org/10.1016/j.tourman.2019.104090 [Accessed 6 September 2025].

Dolnicar, S., 2021. *Market segmentation analysis: understanding it, doing it, and making it useful*. 1st ed. New York: Springer.

Ganta, N. and Krishna, M., 2024. *Price prediction and recommendation of Airbnb property listings using machine learning. SSRN Electronic Journal*. Available at: https://doi.org/10.2139/ssrn.5214669 [Accessed 6 September 2025].

Gibbs, C., Guttentag, D., Gretzel, U., Morton, J. and Goodwill, A., 2018. Pricing in the sharing economy: a hedonic pricing model applied to Airbnb listings. *Journal of Travel & Tourism Marketing*, 35(1), pp.46–56.

Inside Airbnb, 2019. *New York City Airbnb Open Data*. Kaggle. Available at: https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data [Accessed 6 September 2025].

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), pp.651–666.

Kadri, U.F.S.S., 2024. *Machine learning-based categorization of Airbnb listings in NYC. Journal of Soft Computing Paradigm*, 6(3), pp.299–313. Available at: https://doi.org/10.36548/jscp.2024.3.006 [Accessed 6 September 2025].

Liu, M., Hu, S., Ge, Y., Heuvelink, G.B.M., Ren, Z. and Huang, X., 2021. *Using multiple linear regression and random forests to identify spatial poverty determinants in rural China. Spatial Statistics*, 42, 100461. Available at: https://doi.org/10.1016/j.spasta.2020.100461 [Accessed 6 September 2025].

Tang, L.R., Kim, J. and Wang, X., 2024. Forecasting Airbnb demand with machine learning models. *Managerial and Decision Economics*, 45(1), pp.148–160.

Xu, R. and Wunsch, D. (2008). *Clustering*. Hoboken, NJ: Wiley-IEEE Press.