Balancing Tourism Flows with KNN:
A Neighbourhood Busyness-Score Recommender

1. Introduction

Overtourism has become a pressing challenge in major urban centres worldwide (Mihalic, 2020). Major metropolises experience strained infrastructure, resident displacement, and environmental degradation when too many tourists concentrate in specific neighbourhoods (UNWTO, 2018; Santos, Marinheiro & Brito e Abreu, 2024). While policy measures are essential, data-driven recommendations can help redistribute visitor flows organically (Paterlini et al., 2023).

Airbnb pledges to promote more sustainable tourism practices (Airbnb, 2024). One opportunity lies in guiding guests toward underutilised neighbourhoods when their preferred area is at capacity. How can Airbnb address overtourism by recommending less crowded yet similar neighbourhoods within a user's preferred borough?

Using New York City Airbnb Open Data (Dgomonov, 2019), we propose a neighbourhood-level busyness index and a K-Nearest Neighbors (KNN) recommender that suggests quieter but comparable areas. See Bowyer et al. (2025) for the full modelling pipeline.

2. Exploratory Data Analysis

We aggregated all 48,895 listings into 226 neighbourhoods, allowing our model to operate at the neighbourhood level. At the listing stage, 10,237 entries (≈21 %) had `reviews_per_month = NaN`; each also had `number_of_reviews = 0`, so we set those to zero (Soley-Bori, 2013). Minor missing values in `name` and `host_name` were filled with "Unknown," though these fields do not feed our current model. Listings with `price ≥ \$500` were removed before summarisation.

For each neighbourhood, we computed:

- **Price statistics:** neighbourhood-level `price_median` (citywide median = $112, IQR $45–$175), plus borough-level medians (Manhattan $150; Brooklyn $92; Bronx $68; Queens $90; Staten Island $72); right-skewed prices motivated our use of log-transformed summary metrics.

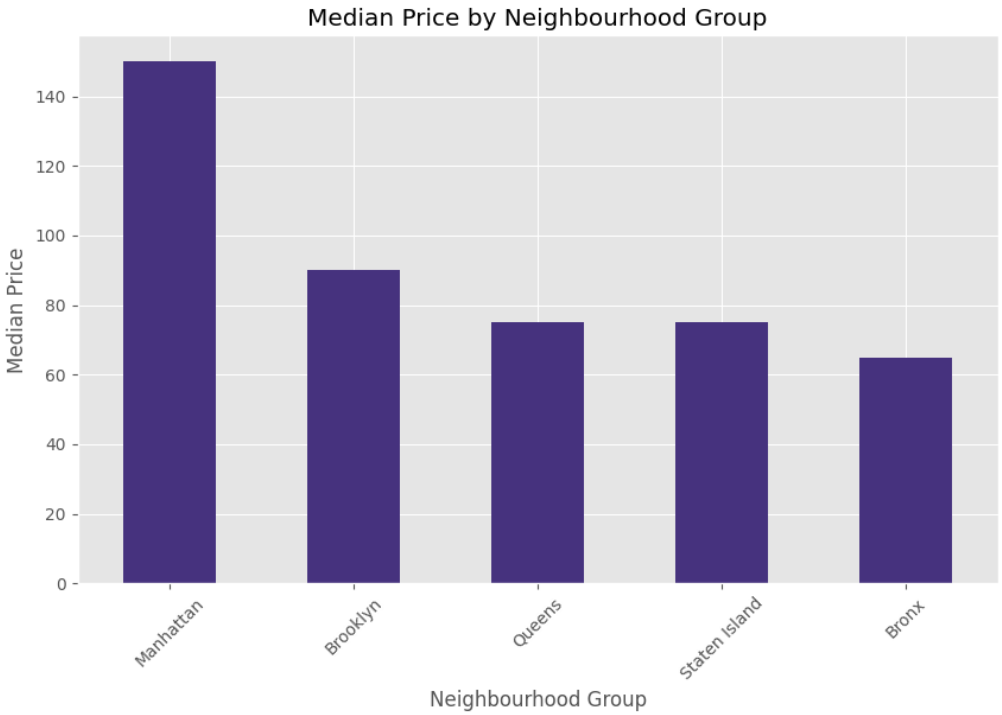


Figure 1. Media Price by Neighbourhood Group

- **Availability and density:** `availability_median` (Manhattan 60 days; Brooklyn 80 days), `listing_count` (e.g., Williamsburg 3,452 listings vs. some Staten Island areas < 20). We also defined `inverse_avg_availability = 1 / (avg_availability_365 + 0.01)`.

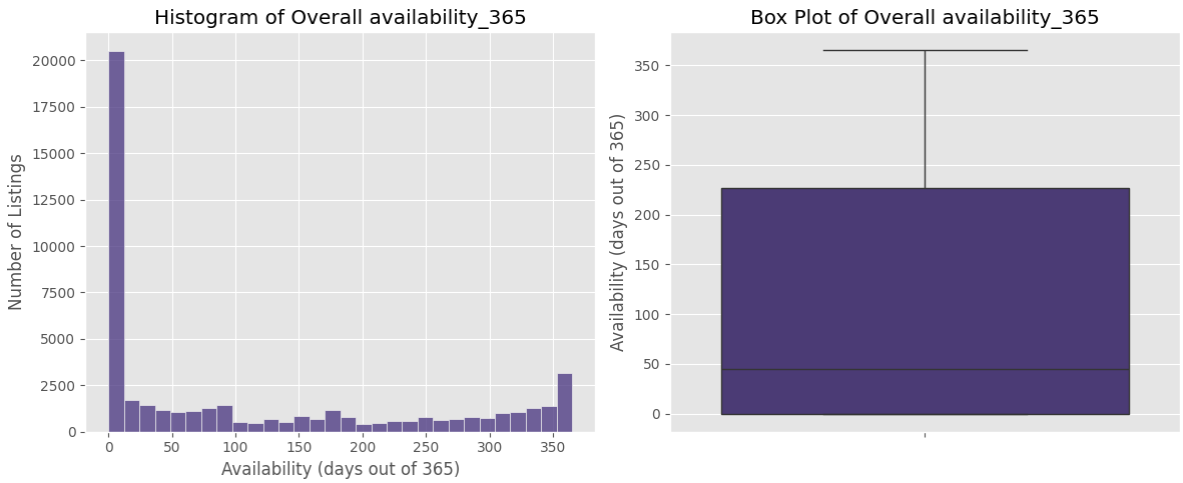


Figure 2. Distribution of Annual Listing Availability in New York City.

- **Review activity:** `avg_reviews_per_month` (most listings < 0.5 after imputation); outer boroughs (Staten Island, Queens, Bronx) averaged higher review rates than Manhattan and Brooklyn, suggesting different local demand dynamics.
- **Room-type mix:** proportions of “Entire home/apt” (~ 52 % citywide), “Private room” (~ 46 %), and “Shared room” (~ 2 %), with considerable neighbourhood-level variation (e.g., Upper East Side ~ 80 % Entire home vs. Bushwick ~ 35 %).

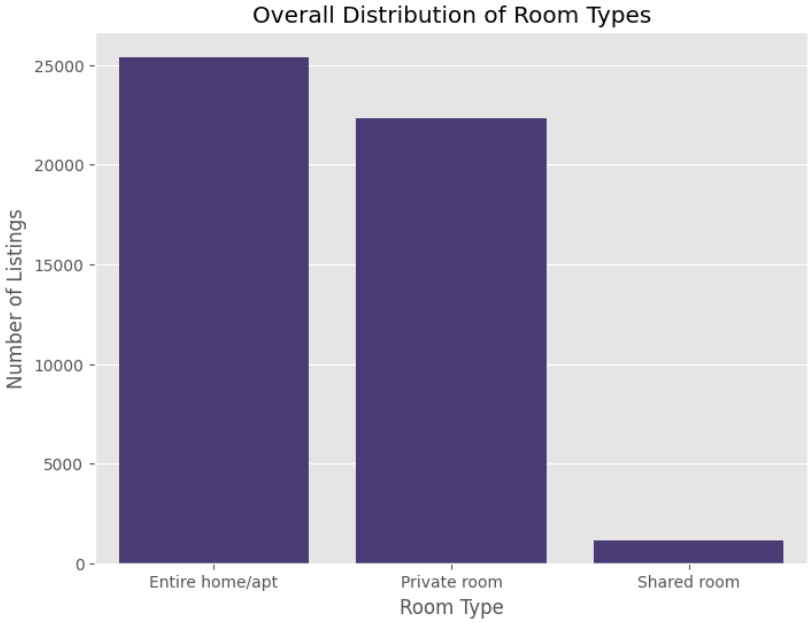


Figure 3a. Room type distribution

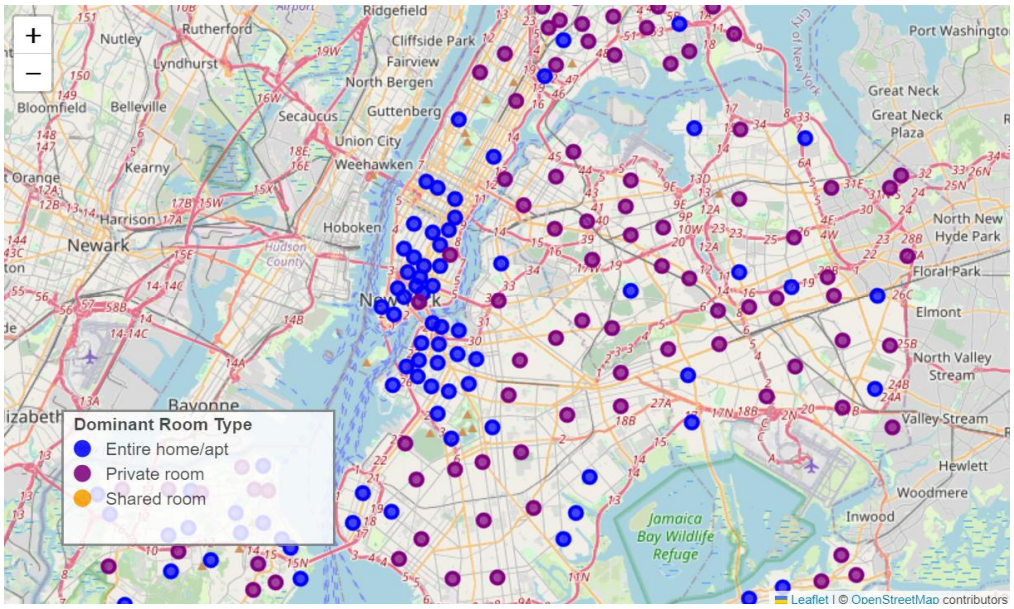


Figure 3b. Folium map to visualise the dominant room type per neighbourhood

- **Minimum-night requirements:** strongly right-skewed (median 3 nights, 75 % at ≤ 5 nights); Manhattan and Brooklyn showed slightly higher averages, indicating pockets of very long minimum stays.

- **Host structure:** proportions of hosts managing exactly 1, 2, 3–5, 6–10, 11–50, or 51+ listings; 66 % of hosts citywide manage one listing, but some neighbourhoods (e.g., Williamsburg) have a long tail of professional hosts, potentially biasing local pricing.
- **Price-tier shares:** proportions in “Budget” (< $69), “Mid-Range” ($69–$174), Premium” ($175–$354), and “Upper Premium” (≥ $355).

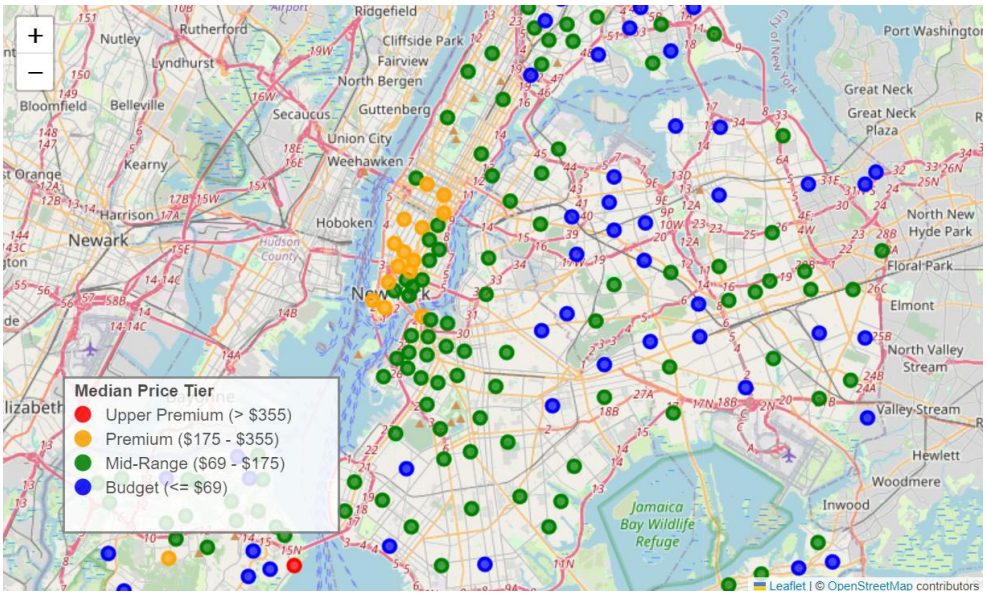


Figure 4. Folium map for neighbourhood price profile

Pairwise correlations among `inverse_avg_availability`, `avg_reviews_per_month`, and `listing_count` were all below |0.30|, confirming that no single metric suffices to capture neighbourhood busyness.

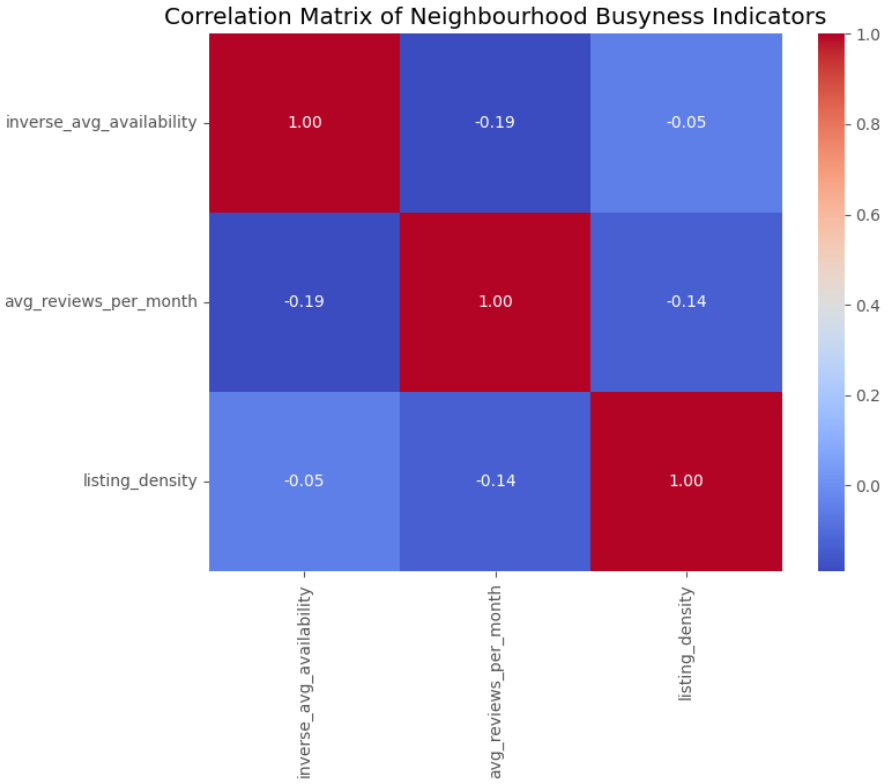


Figure 5. Correlation Matrix of Busyness Indicators

These eighteen MinMax-scaled features form the basis for our composite busyness index and KNN similarity framework, ensuring we account for price, host, review, availability, and room-mix dimensions when comparing neighbourhoods, aligning our recommendations with true market heterogeneity rather than broad borough averages.

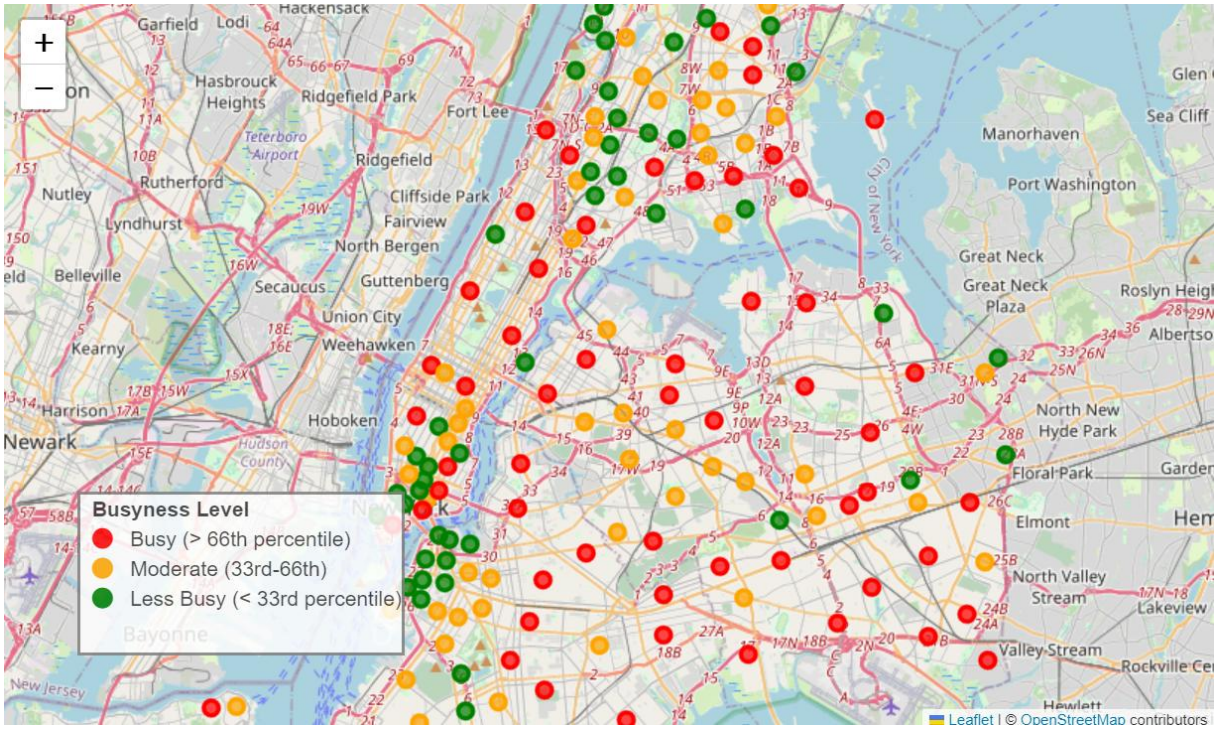


Figure 6. Folium map for neighbourhood busyness

Critical Considerations

Inputting “Unknown” for `name` and `host_name` has little impact here, but would hamper any future text-based or host-reputation models. The 2019 snapshot may not reflect post-COVID shifts in supply and demand (Choi & Kim, 2024). The strong right skewness in minimum-night requirements and price suggests potential outlier influence. Future work should evaluate robust median-based scaling or trimmed means as alternatives to the current approach. Finally, while low correlations among our core busyness inputs justify the use of a composite index, we should validate that this proxy accurately predicts search-cancellation or booking-delay events once integrated into a live environment.

3. Modelling

We built the recommendation engine using K-Nearest Neighbors (KNN), which retrieves similar points without requiring labelled training data or heavy tuning (Cover & Hart, 1967; Hastie, Tibshirani & Friedman, 2009). KNN simply returns existing neighbourhoods whose aggregated profiles most closely match a target.
Each of the 226 neighbourhood profiles comprises our eighteen MinMax-scaled features and a composite busyness_score, computed as the average of three MinMax-scaled inputs:

1. **inverse_avg_availability** = 1 / (avg availability + 0.01)
2. **avg_reviews_per_month**
3. **listing_count**

By construction, a higher busyness_score indicates a more congested neighbourhood. We then fit KNN on these eighteen-dimensional vectors. We initially selected **k = 3** nearest neighbours for exploratory purposes. At query time, the user chooses a borough and neighbourhood; KNN returns the three most similar neighbourhoods overall and separately identifies the single neighbour whose busyness_score is strictly lower, ensuring each suggested alternative is demonstrably less busy. Because KNN uses Euclidean distance, all numeric features (including busyness_score) were scaled to [0, 1].

This proof-of-concept KNN can be integrated into Airbnb's interface to redirect guests from areas at capacity to quieter, yet comparable, neighbourhoods. The model shows promise for balancing demand, reducing local overcrowding, and surfacing underutilised inventory. The full implementation is available on GitHub (Bowyer et al., 2025).

Figure 7. Proof of Concept Widget of the model suggesting a less busy neighbourhood

Critical Considerations

Aggregating listings into neighbourhood profiles smooths over intra-neighbourhood heterogeneity and relies on MinMax scaling, so extreme values can disproportionately influence neighbour rankings. KNN's use of Euclidean distance assumes all eighteen features contribute equally, yet dimensions like price variability versus host structure may merit different weights; we have not performed hyperparameter tuning or distance weighting. Our choice of k = 3 was not cross-validated, leaving potential performance gains unexplored. By design, KNN cannot account for temporal dynamics—our model uses a static 2019 snapshot (Dgomonov, 2019), so sudden demand shifts or seasonality are not captured. Small neighbourhoods with few listings yield unstable feature estimates, meaning

recommendations for those areas may be unreliable. Finally, because we do not validate against real search-cancellation or booking-delay data, we cannot yet confirm that “less busy” suggestions translate to lower user drop-off in practice. Future work should address these concerns through robust validation, feature-weight optimisation, and incorporation of temporal booking patterns.

4. Conclusions

Our neighbourhood-level busyness_score–based recommender offers a practical, data-driven approach for Airbnb to alleviate local overcrowding by directing guests toward quieter but comparable areas. By synthesising availability, review activity, listing density, and composition into a unified neighbourhood profile, the engine surfaces alternatives that match price and stay requirements while reducing congestion.

This proof-of-concept demonstrates how platforms can utilise their data to support more balanced tourism. Integrating these recommendations with broader strategies, such as promoting off-season travel, partnering with local tourism boards, and involving residents in planning, can lead to more sustainable outcomes (Paterlini et al., 2023; Shafiee, 2024). Recent work also emphasises how adaptive governance and innovative tourism systems can balance growth with sustainability (Bisht et al., 2025). Our engine represents a tangible step toward data-driven mitigation of overtourism.

References

Airbnb (2024) *Sustainability and Community Update*. Available at: https://s26.q4cdn.com/656283129/files/doc_downloads/governance_doc_updated/2024/12/Airbnb-2024-Sustainability-and-Community-Update.pdf (Accessed: 25 May 2025).

Bisht, A. *et al.* (2025) 'Managing seasonality and overtourism for sustainable tourism development: challenges and strategies', *International Journal of Research and Review*, 12(3). Available at: https://www.ijrrjournal.com/IJRR_Vol.12_Issue.3_March2025/IJRR55.pdf (Accessed: 1 June 2025).

Bowyer, M. *et al.* (2025) *Airbnb Busyness Recommendation Engine*. Available at: https://github.com/tm3-machine-learning/airbnb (Accessed: 25 May 2025).

Choi, S. and Kim, S. (2024) 'Impact of COVID-19 pandemic on Airbnb listings in New York City: challenges and opportunities for urban housing sustainability', *Sustainability*, 16(20), p. 9140. doi:10.3390/su16209140 (Accessed: 24 May 2025).

Cover, T.M. and Hart, P.E. (1967) 'Nearest Neighbor Pattern Classification', *IEEE Transactions on Information Theory*, 13(1), pp. 21–27. Available at: https://ieeexplore.ieee.org/document/1053964 (Accessed: 2 June 2025).

Dgomonov (2019) *New York City Airbnb Open Data*. Available at: https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data (Accessed: 6 May 2025).

Mihalic, T. (2020) 'Conceptualising overtourism: A sustainability approach', *Annals of Tourism Research*, 84, p. 103025. Available at: https://www.sciencedirect.com/science/article/pii/S0160738320301699 (Accessed: 1 June 2025).

Paterlini, M. *et al.* (2023) 'Can local tours disperse tourists from city centres?', *Journal of Urban Tourism Studies*. Available at: https://www.tandfonline.com/doi/full/10.1080/13683500.2023.2218607 (Accessed: 1 June 2025).

Santos, T.M., Marinheiro, R.N. and Brito e Abreu, F. (2024) 'Wireless crowd detection for smart overtourism mitigation', *Smart Tourism Journal*. Available at: https://arxiv.org/abs/2402.09158 (Accessed: 1 June 2025).

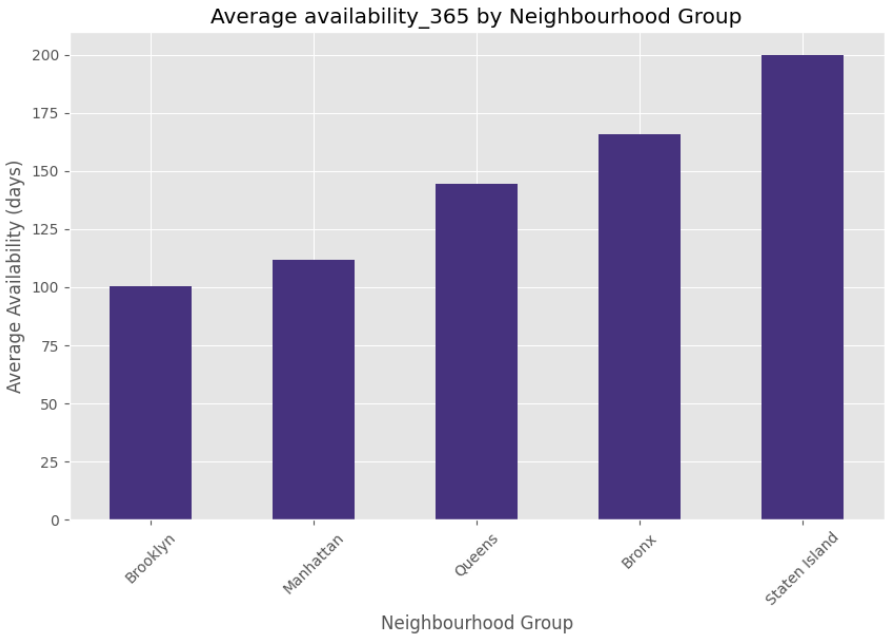
Shafiee, M.M. (2024) 'Navigating overtourism destinations: Leveraging smart tourism solutions for sustainable travel experience', *Smart Tourism*, 5(2), p. 2841. Available at: https://www.researchgate.net/publication/385328329 (Accessed: 3 June 2025).

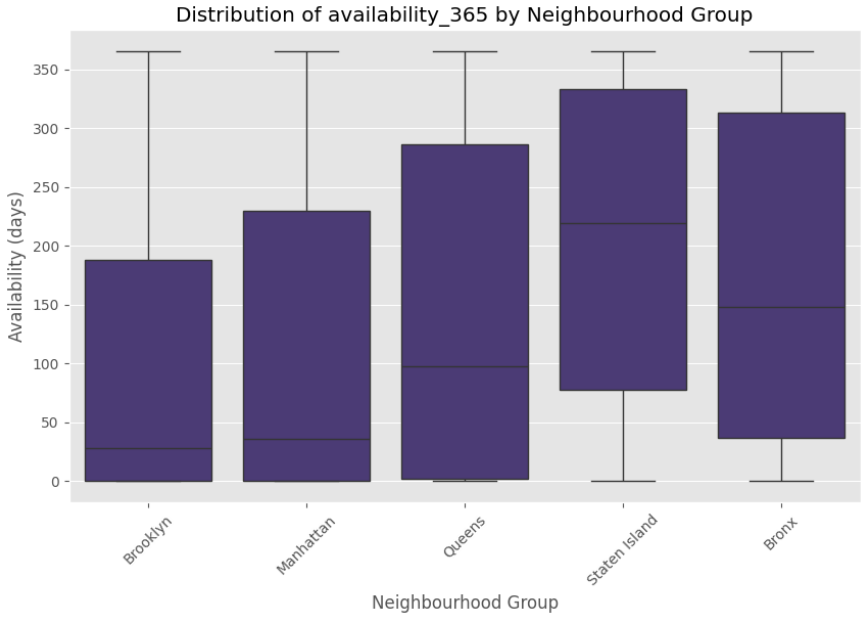
Soley-Bori, M. (2013) *Dealing with Missing Data: Key Assumptions and Methods for Applied Analysis*. Boston University. Available at: https://www.bu.edu/sph/files/2014/05/Marina-tech-report.pdf (Accessed: 3 June 2025).

UNWTO (2018) *Overtourism? Understanding and Managing Urban Tourism Growth Beyond Perceptions*. World Tourism Organization. Available at: https://www.e-unwto.org/doi/book/10.18111/9789284419999 (Accessed: 2 June 2025).

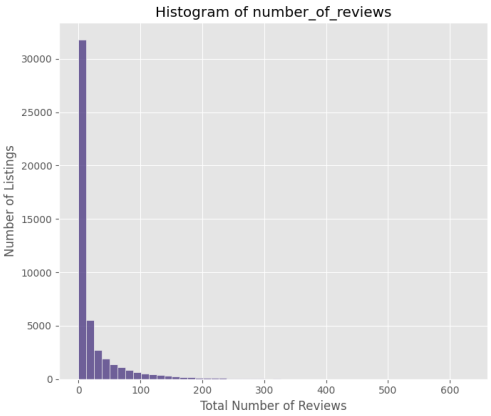
Appendix – Exploratory Data Analysis Visuals.

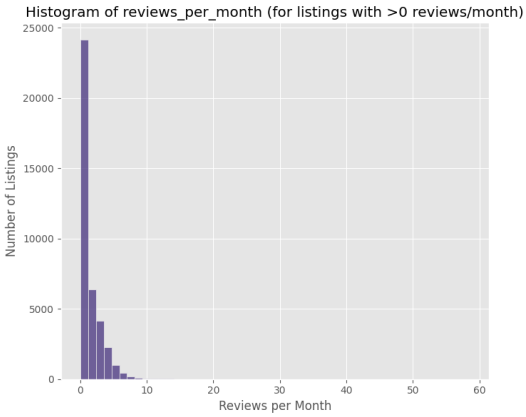
A1: Availability.

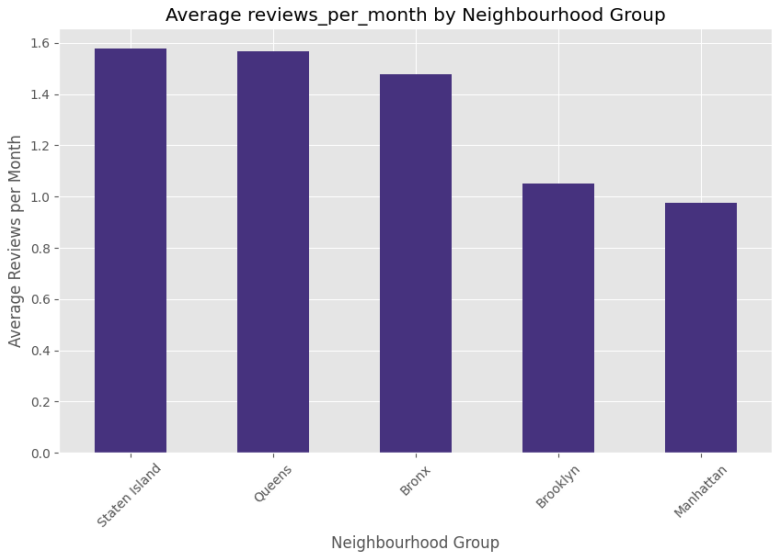


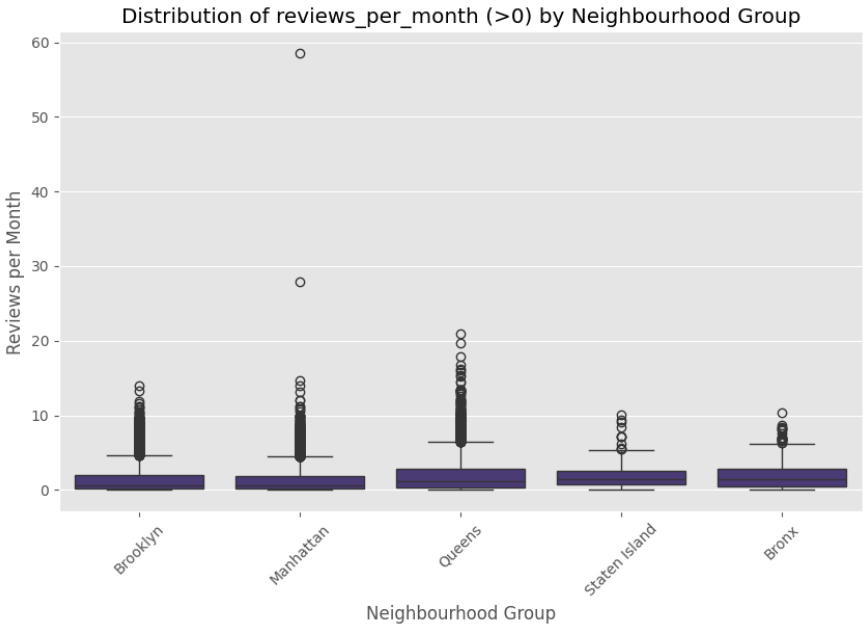


A2: Number of Reviews.

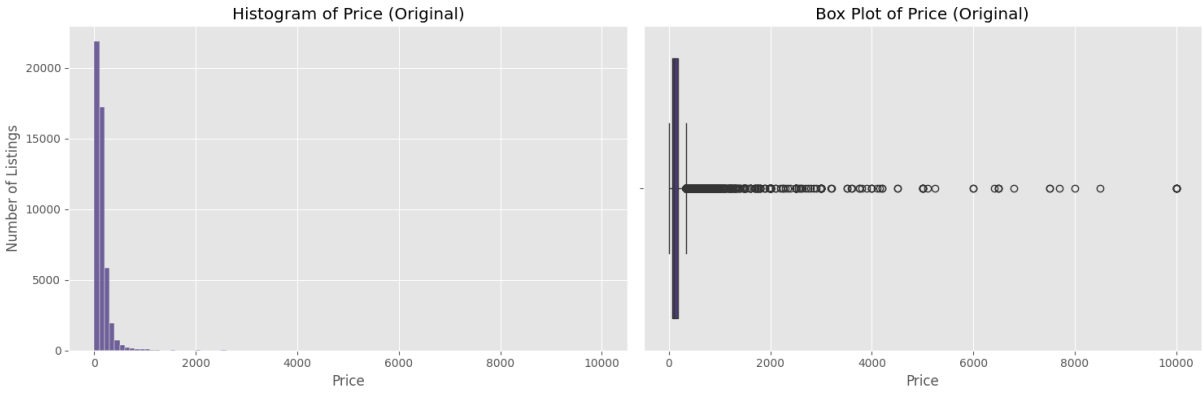


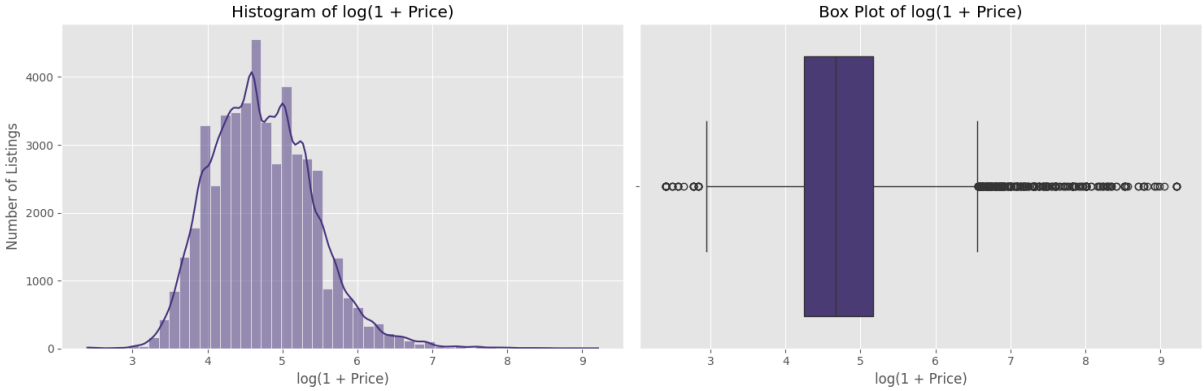


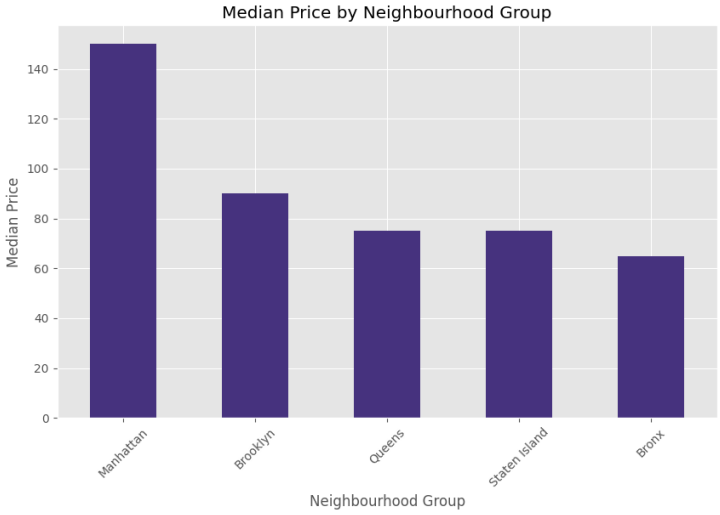


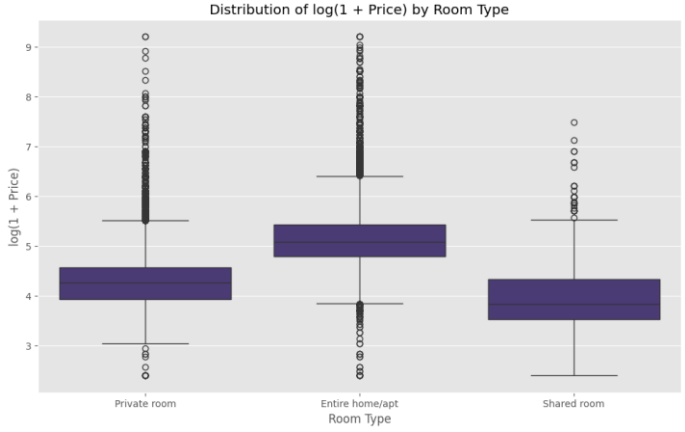


A3: Price Analysis.

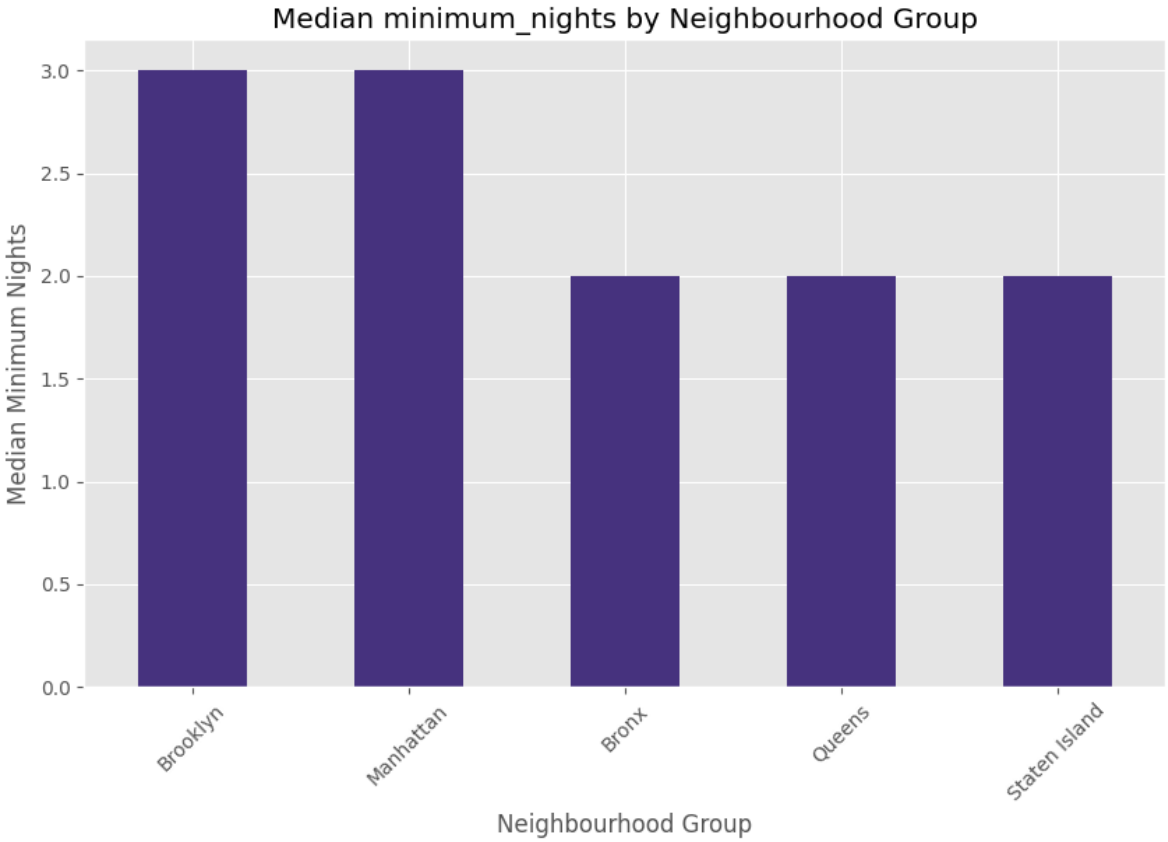




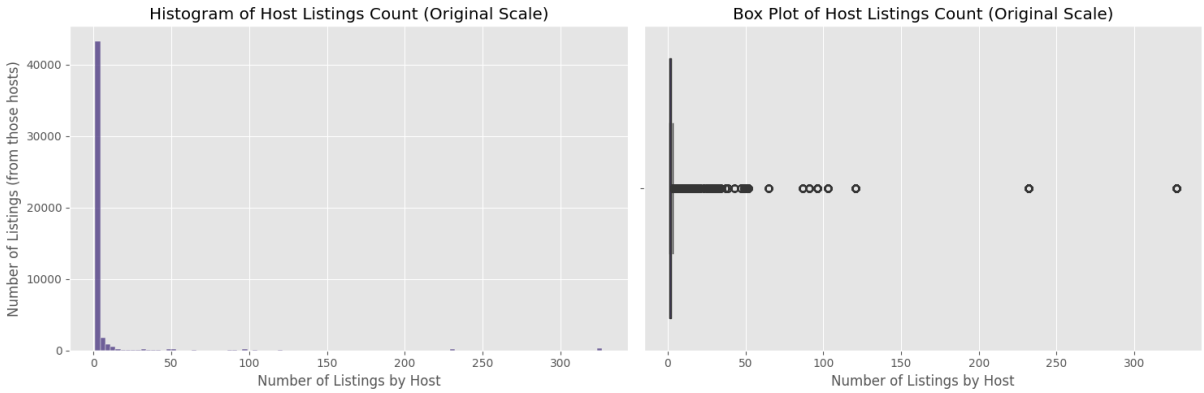


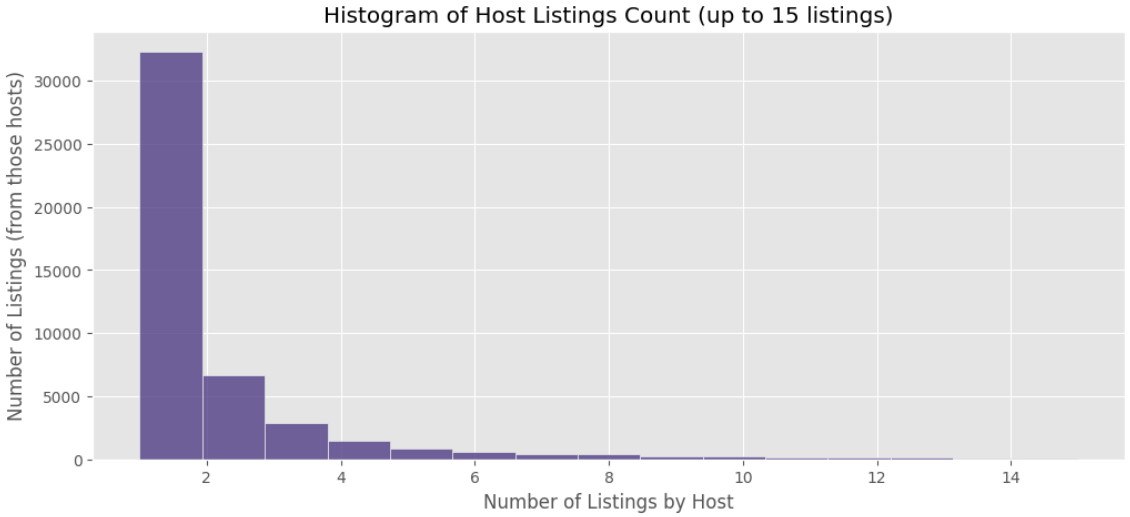


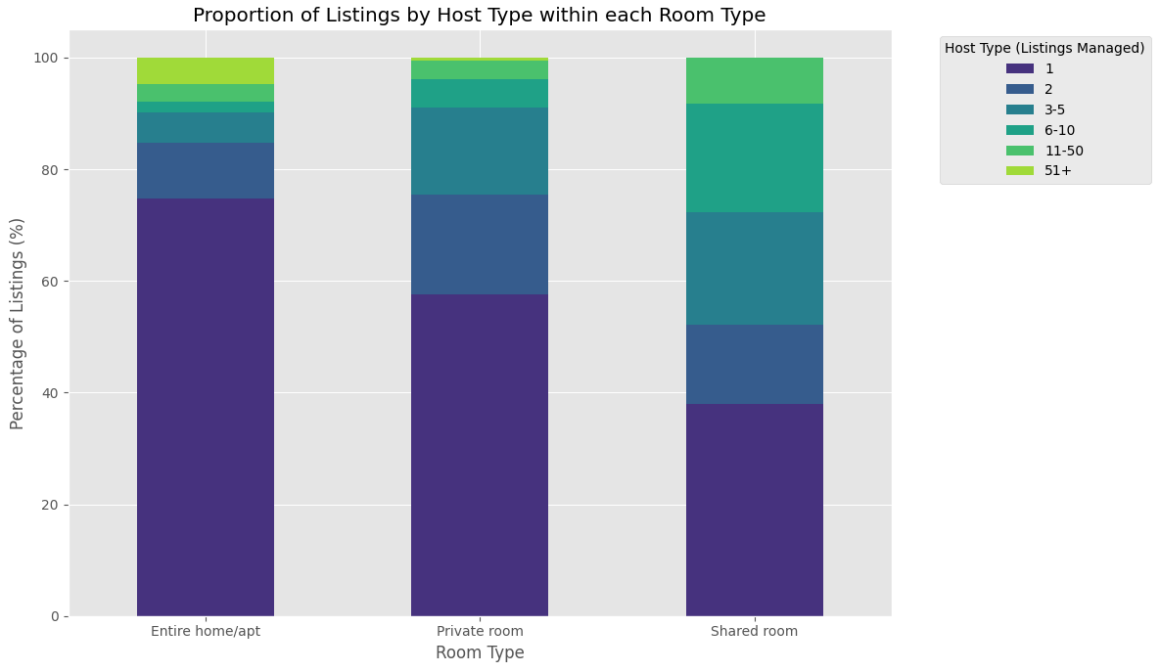
A4: Minimum Nights.



A5: Host Listing Count.







A6: Correlation Matrix of Price and Numerical Features.