

NA_Group2's Group Project

Declaration of Authorship

We, [NA_Group2], pledge our honor that the work presented in this assessment is our own. Where information has been derived from other sources, we confirm that this has been indicated in the work. Where a Large Language Model such as ChatGPT has been used we confirm that we have made its contribution to the final submission clear.

Date: 17/12/2024

Student Numbers: 4

Declaration of Large Language Model Use

In this report, Large Language Models, specifically OpenAI's ChatGPT and Anthropic's Claude, were used as supplementary tools. Their usage was managed to ensure the integrity, originality, and quality of the report. The specific ways in which the LLMs contributed are as follows:

1. **Language Enhancement:** The LLMs were used to improve the grammar, sentence structure, readability, and clarity of the text.
2. **Assistance with Code and Data Analysis:** In sections involving programming or data analysis, the LLMs provided guidance on debugging, refining code, optimizing algorithms, and suggesting alternative approaches to analytical problems.
3. **Concept Clarification:** For technical or theoretical components, the LLMs served as a resource to clarify complex concepts, provide background explanations, and ensure accuracy in the presentation of ideas.

Word Count

Text Content: 1,727

Visuals: 6 figures \times 150 per figure = 900

Total: 2,627

| What Went Well | What Was Challenging |
|----------------|----------------------|
|----------------|----------------------|

Brief Group Reflection

| What Went Well | What Was Challenging |
|---|--|
| <p>A clear focus on the 90-day policy as the core theme.</p> <p>Initial observations revealed significant policy violations.</p> <p>Spatial clustering analysis highlighted policy-violating hotspots.</p> <p>Iterative refinement of models improved insights (GWLR $R^2 = 0.55\text{--}0.96$).</p> <p>Identified policy violations' impact on rent and vacancy rates.</p> <p>Combined four quarters of snapshot data to mitigate time-sensitivity limitations.</p> | <p>InsideAirbnb lacks direct data on annual rental days, requiring estimation.</p> <p>Snapshot data is time-sensitive and may not fully capture temporal changes.</p> <p>The logistic regression model had low explanatory power ($R^2 = 0.05$).</p> <p>Residual analysis and Moran's I indicated spatial heterogeneity.</p> <p>Incorporating new variables (e.g., attraction density) was computationally intensive.</p> <p>Assessing the broader socioeconomic impact was limited by data availability.</p> |

Priorities for Feedback

Code Consistency: When working together, our code and methods often feel disconnected and lack coherence. How can we ensure clearer, more consistent approaches across the team?

Accuracy of Policy Violation Estimates: Given the limitations of InsideAirbnb data (e.g., no direct measure for annual rental days), we used proxies like reviews and minimum stay to estimate violations. We'd appreciate guidance on how to validate or improve this estimation method to reduce bias.

Model Discrepancy and Improvement: While GWLR significantly improved explanatory power ($R^2 = 0.55\text{--}0.96$), the initial logistic regression model had very low explanatory power ($R^2 = 0.05$). We would appreciate feedback on whether there are key factors missing in the initial model and how to avoid potential overfitting or over-reliance on local weights in the GWLR model.

1. Who collected the InsideAirbnb data?

The data was collected, integrated, and published by Inside Airbnb's policy and housing researchers, led by founder Murray Cox, an activist using data-driven insights to address housing challenges (Airbnb, 2024).

2. Why did they collect the InsideAirbnb data?

Inside Airbnb collects data to help communities understand Airbnb's impact on housing and neighborhoods. They aim to provide data that empowers communities to make informed decisions, manage short-term rentals, protect residents from negative effects like rising rents and housing shortages, and support collective efforts by residents and activists (Cox, 2023).

3. How did they collect it?

Inside Airbnb used Python-based web scraping to collect publicly available data from the Airbnb website (Alsudais, 2021). The data is captured as a time-specific snapshot, including availability calendars (up to 365 days), reviews, names, photos, and other publicly visible details. After collection, the data was cleaned, analyzed, and aggregated. Airbnb's categorization of unavailable nights (booked or blocked) and location anonymization were retained unchanged, while neighborhood names were refined using city-defined boundaries for accuracy (Airbnb, 2024).

4. How does the method of collection (Q3) impact the completeness and/or accuracy of the InsideAirbnb data? How well does it represent the process it seeks to study, and what wider issues does this raise?

The Inside Airbnb data reveals key market trends and geographic distribution of listings, offering a foundation for analyzing short-term rental impacts on residential neighborhoods. But it has the following limitations:

1. The data includes only Airbnb listings, which may be removed or hidden by Airbnb due to business or regulatory pressures. For example, in 2015, Airbnb deleted over 1,000 entire-home listings in New York (Cox and Slee, 2016).
2. The data's snapshot nature cannot reflect ongoing changes, such as newly added or removed listings. Differences between dataset versions can also affect reliability and reproducibility if the version used is not specified.
3. The data excludes off-platform bookings, such as private arrangements to avoid platform fees or bypass rental day regulations. Additionally, booked dates are marked as unavailable, creating the false impression that popular listings are unrentable. These factors lead to an underestimation of actual bookings and occupancy rates.
4. Geolocation data is anonymized, shifting locations by up to 150 meters and scattering listings within buildings, which reduces accuracy for neighborhood-level analysis.

5. Airbnb sometimes assigns the same ID to both ‘Experiences’ and ‘Listings,’ causing ‘Experience’ reviews to be misclassified as ‘Listing’ reviews and leading to inaccuracies in review data (Alsudais, 2021).
-

5. What ethical considerations does the use of the InsideAirbnb data raise?

1. Preventing Misuse of Data for Commercial Gain

Inside Airbnb’s data is meant to reveal the impacts of short-term rentals and to support local communities. However, malicious exploitation of the data—such as investors using it to identify high-profit areas for short-term rentals—could lead to negative consequences, including increased property purchases for rentals, displacement of residents, and worsening housing crises. This undermines the platform’s mission and further harms vulnerable groups. Researchers and policymakers have a responsibility to ensure that their use of the data aligns with community interests and to avoid sharing data in ways that enable harmful commercial use.

2. Protecting User Privacy

Since the data is collected through web scraping, Airbnb hosts and guests may not have given consent for its use. Even anonymized geographic data can unintentionally expose sensitive information, especially in small communities or for standalone properties. Publicly identifying short-term rental activity could subject hosts and guests to harassment, discrimination, or financial loss. Research outputs should generalize findings to avoid unintentionally identifying individuals and use neutral language to prevent stigmatization of particular groups.

3. Ensuring Representation of Vulnerable Communities: Promoting Equity in Data Interpretation

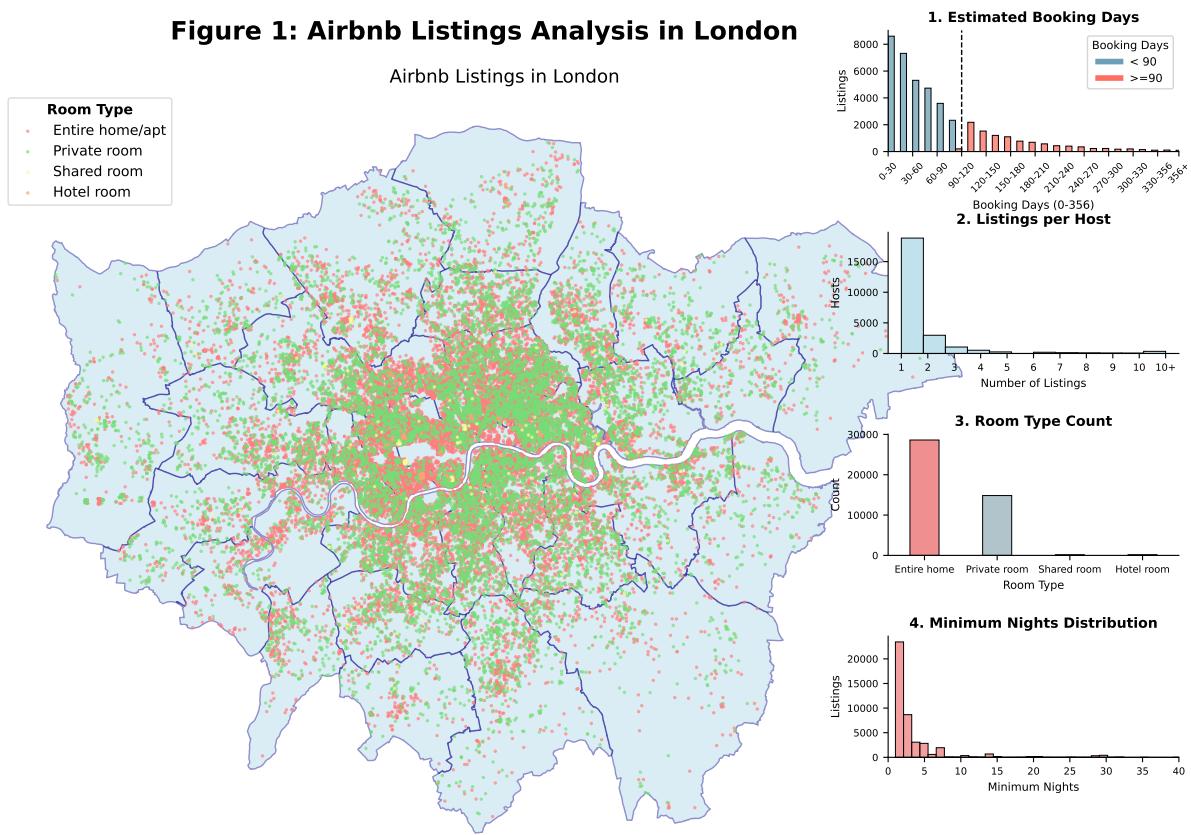
While Inside Airbnb data is openly available, the ability to analyze and interpret it is often limited to technically skilled individuals, such as researchers or local activists. This can result in policies that prioritize the perspectives of these groups while neglecting the needs of underrepresented or vulnerable communities. To address this imbalance, researchers and policymakers should engage marginalized communities in the interpretation process, present findings in accessible formats, and ensure data-driven decisions promote social equity, as highlighted in Data Feminism (D’Ignazio and Klein, 2020).

6. With reference to the InsideAirbnb data (i.e. using numbers, figures, maps, and descriptive statistics), what does an analysis of Hosts and the types of properties that they list suggest about the nature of Airbnb lettings in London?

The 90-Day Rule and Airbnb Commercialization

Airbnb has reduced long-term rentals, raised rents, and driven commercialization in London’s housing market. The 90-day rule limits entire-home short-term rentals to 90 days per year without special planning permission (Greater London Authority, 2023), but enforcement depends on self-reporting, which is weak. To address this, we use InsideAirbnb data to analyze booking patterns, host behavior, and property types to assess commercialization and compliance risks. Let’s start by conducting an initial observation of the data:

Figure 1: Airbnb Listings Analysis in London



Key Findings

1. Spatial Hotspot:

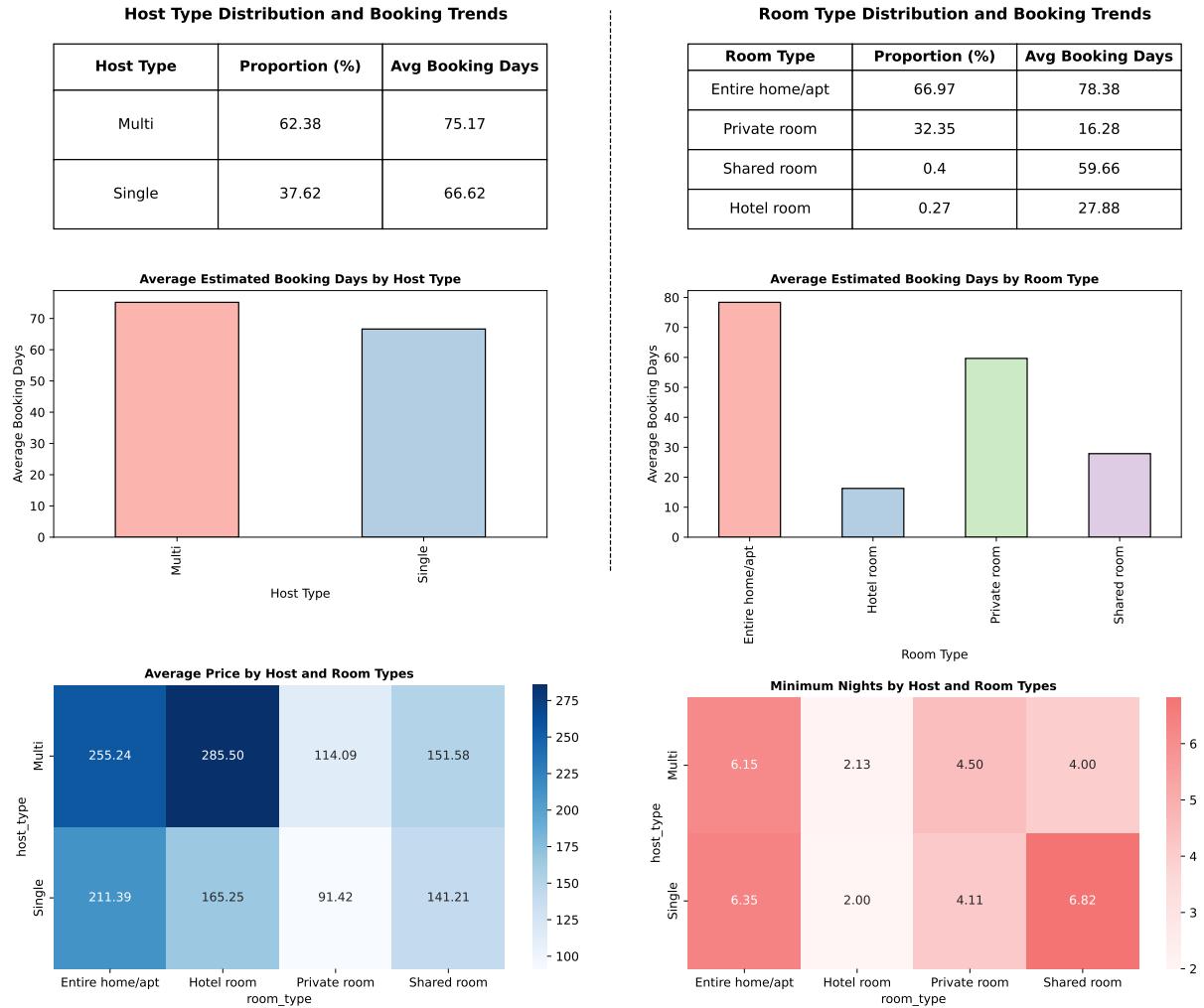
Central London has the highest density of short-term rentals, with entire homes dominating high-demand, high-violation areas of the 90-day rule (Figure 1, Map).

2. Frequent Breaches of the 90-Day Limit:

Booking data, based on reviews, pricing, and minimum nights (InsideAirbnb, 2023), shows entire homes and hotel-like properties often exceed the 90-day limit, especially in commercialized areas (Figure 1, Panel 1).

To address the limitations of using only snapshot data, we merged datasets from December 2023 to September 2024 and removed duplicates to ensure greater accuracy:

Figure 2: Analysis of Airbnb Listings by Host and Room Types



3. Multi-Listing Hosts Dominate the Market:

Multi-listing hosts manage 62.38% of Airbnb listings with longer average booking durations than single-listing. They primarily operate entire and hotel-like properties, which are highly commercialized with shorter stays and higher prices (Figure 2, Top Left; heatmap).

4. Entire Homes are the Highest Risk:

Entire homes (66.97% of listings) have the highest average booking days and are most likely to breach the 90-day rule, followed by hotel-like properties with short minimum stays (2.13 nights) and high commercialization (Figure 2, Top Right; heatmap).

Initial conclusion

InsideAirbnb data reveals London's highly commercialized Airbnb market with significant 90-day rule violations, concentrated in high-risk areas. Multi-listing hosts and commercialized properties are key drivers, requiring further spatial and regression analysis to assess impacts and improve enforcement.

7. Drawing on your previous answers, and supporting your response with evidence (e.g. figures, maps, EDA/ESDA, and simple statistical analysis/models drawing on experience from, e.g., CASA0007), how could the InsideAirbnb data set be used to inform the regulation of Short-Term Lets (STL) in London?

We decided to further analyze the December 2023 Inside Airbnb dataset to provide evidence on the link between Airbnb activity and policy violations. The next steps are:

1. Identifying Violation Hotspots and Impacts:

Spatial clustering will identify areas with high concentrations of 90-day rule violations. These hotspots will be analyzed for impacts on local communities, focusing on rising rents and reduced housing availability—issues tied to Airbnb-driven gentrification ([smith2006?](#)).

2. Analyzing Drivers of Violations:

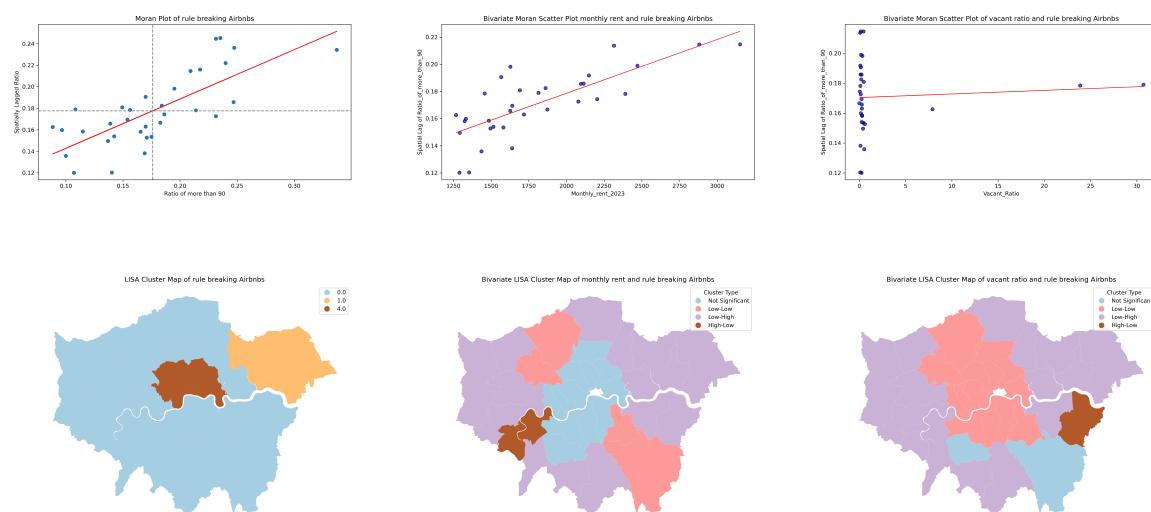
Regression analysis helps identify key factors driving these violations (e.g. property type, host behavior, price, location...)

This approach will highlight the need for effective regulation, and clarify enforcement priorities to enhance the efficiency of enforcement efforts.

Spatial Analysis of Policy Violations / Local Impacts

1. Hotspot Identification: We analyzed the spatial clustering of Airbnb rule-breaking properties using Moran's I and LISA cluster maps (Figure 3). High-High clusters were found in central boroughs, such as Westminster, and eastern areas like Hackney, where violations are linked to a combination of high tourism demand, profitability of short-term rentals, and housing market pressures (Bosma and Doorn, 2024).

Figure 3: Results of Moran and LISA analysis of rule breaking Airbnbs, monthly rent and vacancy ratio

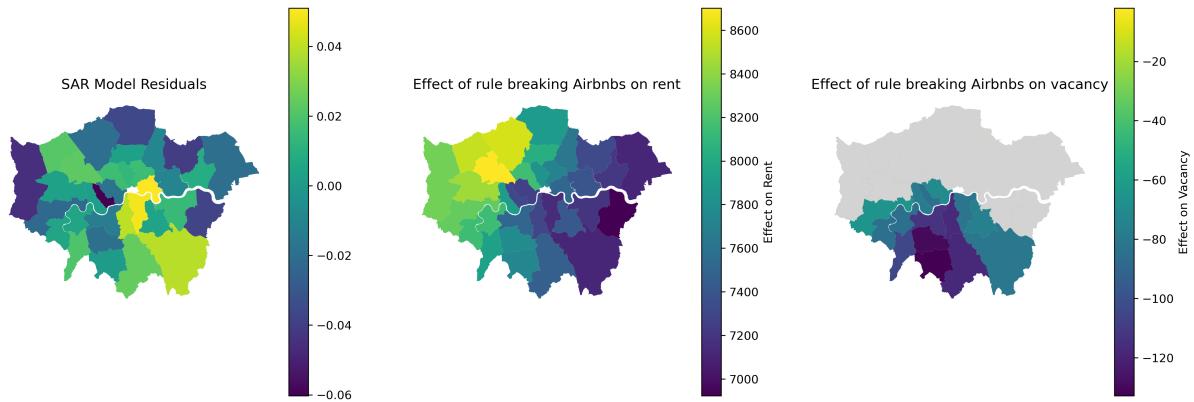


2. Housing Market Impacts

To quantify these impacts, we applied SAR and GWR models (Figure 4). The SAR analysis showed that violations contributed to rising rents and increased vacancy rates, with the

strongest effects observed in central areas where tourism dominates and in eastern boroughs with emerging rental markets. GWR results highlighted spatial variability, aligning with findings from (Jain *et al.*, 2021), the highest rent surges in central London and higher vacancy rates in eastern boroughs.

Figure 4: Results of SAR and GWR Analysis of the Effect of Rule-Breaking Airbnbs on Monthly Rent and Vacancy Ratio

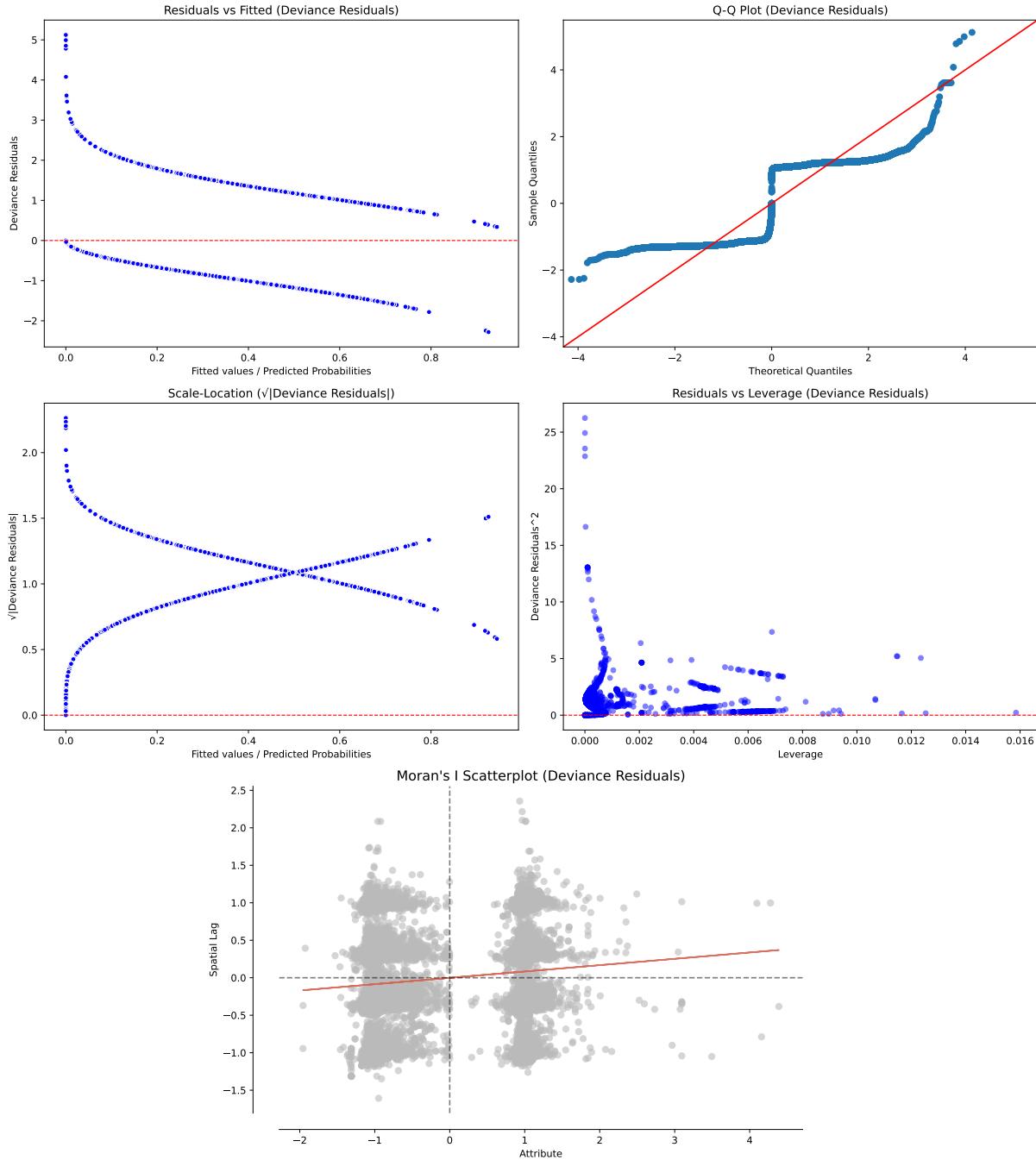


From the analysis above, enforcing the 90-day policy is essential to address rising rents, increasing vacancy rates, and spatial inequality driven by short-term rentals, thereby preserving housing affordability and stability in affected hotspots.

Analysis of Factors Associated with Policy-Violating Listings

1. **Logistic Regression:** To explore the relationship in the study area between variables and their spatial distribution characteristics, we adopted a logistic regression model to establish the model and evaluated the performance of the model through the residual and spatial autocorrelation test.

Figure 5. Visualisation of Residuals and Moran's I



Key property and host characteristics were identified as significant drivers of policy violations ($p < 0.001$). However, the low pseudo R-squared (0.05) suggests the model explains only a small portion of the variability in policy violations.

Residual analysis and Moran's I test ($I = 0.0855$, $p = 0.001$) highlighted weak but significant spatial patterns, indicating **missing factors** and **spatial heterogeneity**. Further model improvements are needed to better capture these complexities.

-
2. **Geographically Weighted Regression:** To address spatial heterogeneity, research suggests adopting the Geographically Weighted Regression (GWR) model, which allows variable coefficients to vary by location. Meanwhile, previous studies have shown a strong association between public transport density, green space density, and tourist attraction density with Airbnb distribution (Xu *et al.*, 2020). Thus, we incorporate these

variables into the model to examine whether they similarly influence the occurrence of non-compliant listings. The data for these variables were obtained from OpenStreetMap (Geofabrik GmbH, 2024).

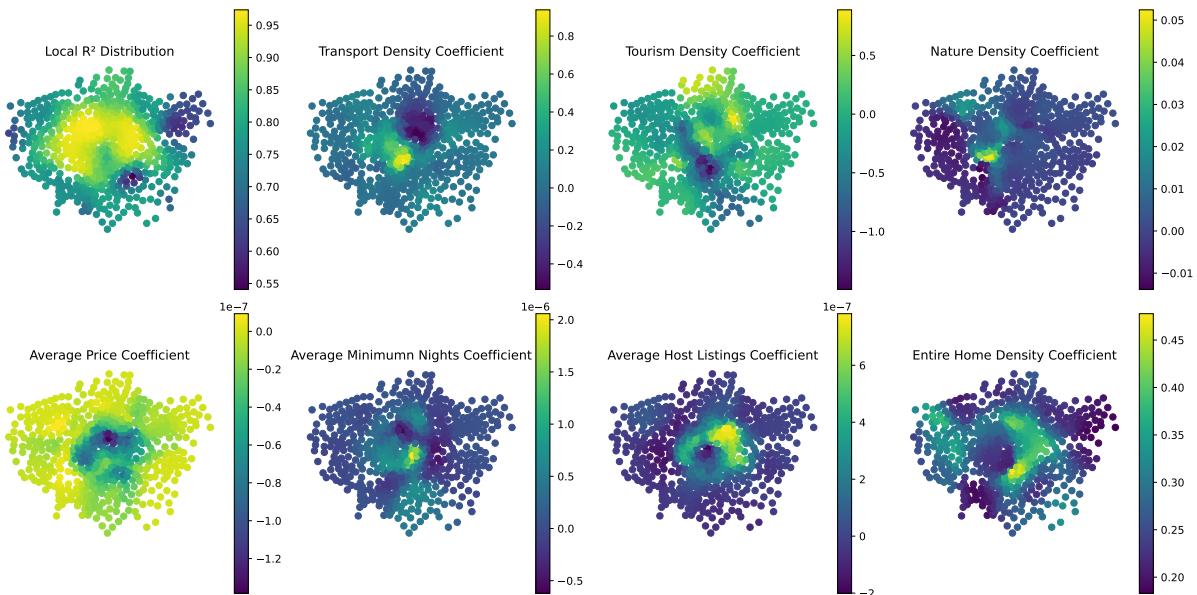


Figure 6: Spatial Distribution of GWR Results for Airbnb Listings Analysis

The GWR model demonstrates strong explanatory power, accounting for over 80% of the variation in non-compliant listings across London ($R^2 \geq 0.8$) and 95% in concentrated central areas ($R^2 = 0.95$). Key factors driving non-compliance include public transport density and tourist attraction density.

However, significant spatial heterogeneity exists. In the lower-left central region, public transport density shows a strong positive correlation, while the upper-right shows a moderate negative correlation. For tourist attraction density, the opposite pattern is observed: a strong negative correlation in the lower-left and a moderate positive correlation in the upper-right. This suggests that no single factor uniformly explains non-compliance across all areas.

Conclusion / Policy Recommendations

We conducted a spatial analysis using Inside Airbnb data and found that illegal listings in London exhibit a clear clustering pattern. These listings significantly impact local housing market rental levels and vacancy rates, highlighting the necessity and reasonableness of the current 90-day short-term rental policy.

Further multi-factor analysis reveals that while illegal listings show some correlation with factors such as host type, property type, minimum rental period, and price, their overall impact is limited. Instead, traffic network density and tourist density emerge as the key drivers of illegal listing distribution. However, their influence varies significantly across regions, indicating that a single regulatory policy cannot effectively address illegal activities in all areas.

Policy Recommendations 1. Citywide Housing Database and Data Sharing: We recommend that the government establish a citywide housing database and require short-term rental platforms to sign data-sharing agreements. These platforms should regularly upload housing information and usage data to support regulatory work.

2. **Learning from European Cities:** Drawing on the experience of cities like Barcelona, short-term rental platforms should be compelled to remove unregistered listings and face stricter penalties for violations. (Bei, 2025)
3. **Differential Regional Regulation:** Building on London's Housing Act 2016 (UK Government, 2016), we suggest implementing region-specific regulations. Based on data analysis, enforcement resources can be allocated more effectively to hotspot areas of illegal listings, optimizing patrols and rectification measures.

These measures would improve regulatory efficiency and help reduce illegal activities across London's diverse regions.

References

- Airbnb, I. (2024) 'Inside airbnb'. Available at: <https://insideairbnb.com/>.
- Alsudais, A. (2021) 'Incorrect data in the widely used Inside Airbnb dataset', *Decision Support Systems*, 141, p. 113453. Available at: <https://doi.org/10.1016/j.dss.2020.113453>.
- Bei, G. (2025) 'The spatial effect of short-term rental regulations: The comparison between barcelona and paris', *Cities*, 158, p. 105603. Available at: <https://doi.org/10.1016/j.cities.2024.105603>.
- Bosma, J.R. and Doorn, N. van (2024) 'The gentrification of airbnb: Closing rent gaps through the professionalization of hosting', *Space and Culture*, 27(1), pp. 31–47. Available at: <https://journals.sagepub.com/doi/full/10.1177/12063312221090606>.
- Cox, M. (2023) *Inside Airbnb: Dallas*. Inside Airbnb. Available at: <https://insideairbnb.com/reports/inside-airbnb-dallas-march-2023.pdf>.
- Cox, M. and Slee, T. (2016) *How Airbnb's data hid the facts in New York City*. Inside Airbnb. Available at: <https://insideairbnb.com/reports/how-airbnbs-data-hid-the-facts-in-new-york-city.pdf>.
- D'Ignazio, C. and Klein, L.F. (2020) *Data Feminism*. Cambridge: The MIT Press (Strong Ideas).
- Geofabrik GmbH (2024) 'Greater london OSM data download'. Available at: <https://download.geofabrik.de/europe/united-kingdom/england/greater-london.html>.
- Greater London Authority (2023) 'Guidance on short-term and holiday lets in london'. <https://www.london.gov.uk>.
- InsideAirbnb (2023) 'Inside airbnb: Adding data to the debate'. <http://insideairbnb.com>.
- Jain, S. et al. (2021) 'Nowcasting gentrification using airbnb data', *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), pp. 1–21. Available at: <https://dl.acm.org/doi/10.1145/3449112>.
- UK Government (2016) 'Housing and planning act 2016, c. 22 (UK)'. Available at: <https://www.legislation.gov.uk/ukpga/2016/22/contents>.

Xu, F. *et al.* (2020) 'The influence of neighbourhood environment on Airbnb: A geographically weighed regression analysis', *Tourism Geographies* [Preprint]. Available at: <https://www.tandfonline.com/doi/abs/10.1080/14616688.2019.1586987> (Accessed: 15 December 2024).