DeskB's Group Project

Declaration of Authorship

We, [**DeskB**], confirm 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: 19th December 2023

Student Numbers: 20017359 23032922 23081403 23103585 23130397

Brief Group Reflection

What Went Well	What Was Challenging
data description	plotting
data cleaning	SVM classifier model

Priorities for Feedback

Are there any areas on which you would appreciate more detailed feedback if we're able to offer it?

Frankly, we've encountered lots of confusion towards the topic of this assessment. Especially in the topic selection, among all the predictive topis in the website, we can not propose the very specific quesiton and structure at the beginning. How to build the bridge between NLP recommending system for branding and inform valuable proposal for STL regulation could be the key issue for us.

So, if convinient, we would like to know did we structure the whole report with a solid logical chain. Also, did we successfully propose some constructive and feasible suggestions? And what should be NLP analysis used for proposal looked like in a real company project?

Response to Questions

1. Who collected the data?

The dataset was collected by Murray Cox through automatic scraping from the Airbnb website, specifically for the Inside Airbnb project.

2. Why did they collect it?

The Inside Airbnb project aims to provide an independent perspective, helping the public, researchers, and policymakers understand how Airbnb affects urban housing affordability and community dynamics. It offers insights for policy discussions and social understanding of Airbnb's role in urban environments.

3. How was the data collected?

listings.csv : Inside Airbnb collects its data primarily by scraping information from the Airbnb website. This process involves the following steps:

- Web Scraping: Inside Airbnb employs scripts to rapidly and extensively extract Airbnb listing data, imitating human browsing.
- Data Extraction: Information about each listing, such as location, price, availability and host details, is extracted and compiled.
- Data Aggregation: Aggregated data forms a database for analyzing Airbnb trends and insights across cities and regions.
- Regular Updates: The scraping process is repeated periodically to keep the database current, capturing new listings and updates to existing ones.

4. How does the method of collection impact the completeness and/or accuracy of its representation of the process it seeks to study, and what wider issues does this raise?

The dataset is mostly obtained by scraping information from the Airbnb website, so its breadth and depth of information publicly available on the site may be limited. For instance, detailed information about certain listings might not be fully disclosed, or website terms might restrict access to some data. Moreover, legal and ethical considerations in web scraping, such as data privacy and usage rights, may affect the integrity and accuracy of the data. The content of the website is constantly changing dynamically, but data scraping occurs at intervals, which means the data might not be updated in-real time, potentially leading to information gaps(Prentice and Pawlicz, 2023).

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

5.1 Privacy issues

Whether the dataset has the consent of the owner to disclose its information, e.g., house location, name. Geocoded data is privacy-sensitive and highly likely to expose personal privacy when used to study demographic patterns and behaviours(Bemt *et al.*, 2018). Therefore, It is crucial to obtain the consent of the owners to ensure that their privacy is not infringed upon.

5.2 Legal compliance

Usage of the dataset should comply with laws and regulations such as GDPR, DPA and EDPS. The EDPS 2015 report states that it is not enough to comply with the law in today's digital environment; We must consider the ethical dimensions of data processing(Hasselbalch, 2019). Legal compliance and ethical considerations should be closely combined in the digital age.

5.3 Social responsibility

It is critical to use the dataset correctly, as exposing certain data may result in inequity and bias. The Fairness and Openness Report(Walker and Moran, 2019) emphasizes how to use information responsibly and ethically, as well as the importance to resist the labelling of low-income communities, race, etc. For example, a significant gap in housing prices between different neighbourhoods may reflect economic differences, which may affect perceptions of the social status of those areas. To avoid unwanted consequences, it is necessary to examine how to disclose the tagged attributes of the data.

5.4 Data security

Some sensitive information in the dataset, such as personal descriptions and geographic coordinates, must be stored securely to prevent unauthorized access and misuse. By adjusting the norms of network data use, it is possible to effectively guarantee data security and increase companies' ethical behavior level when processing data(Culnan and Williams, 2009). Thus, attention to data security can prevent unscrupulous individuals from collecting housing data for profit or monitoring purposes.

6. With reference to the data (*i.e.* using numbers, figures, maps, and descriptive statistics), what does an analysis of Hosts and Listing types suggest about the nature of Airbnb lets in London?

6.1 Why should we choose the textual information?

Many studies have analyzed various aspects of Airbnb listings, including price(Zhang *et al.*, 2017), spatial distribution(La *et al.*, 2021), room type(Voltes-Dorta and Sánchez-Medina, 2020), etc. However, the "textual description", with

more impressive potential than numeric fields, also plays a crucial role in shaping renters' first impressions of the listings, contributing to facilitating successful rental transactions. Therefore, we scrutinize the textual features/characteristics from the data, generalize, classify and summarize insightful conclusion which is correlated with the branding potential value(Ji, Li and Yang, 2021).

6.2 What can we dig from the textual information?

Datasets consists of two textual fields: 'Description' and 'Amenities' from the host's self-promotion. 'description' column is to describe advantages and characteristics. 'Amenities' is about facilities affiliated with the listing.

After some cleaning and preprocessing, there are two set of questions corresponding to the two columns respectively.

6.2.1 Which topics would host like to focus on when promoting their properties?

We could use the LDA model to generalize and extract topics to get the most frequent keywords in those topics. After calculating iteratively the model, we determine the best topics' number for summarizing 'descriptions' column should be 16. (*Figure1a*)

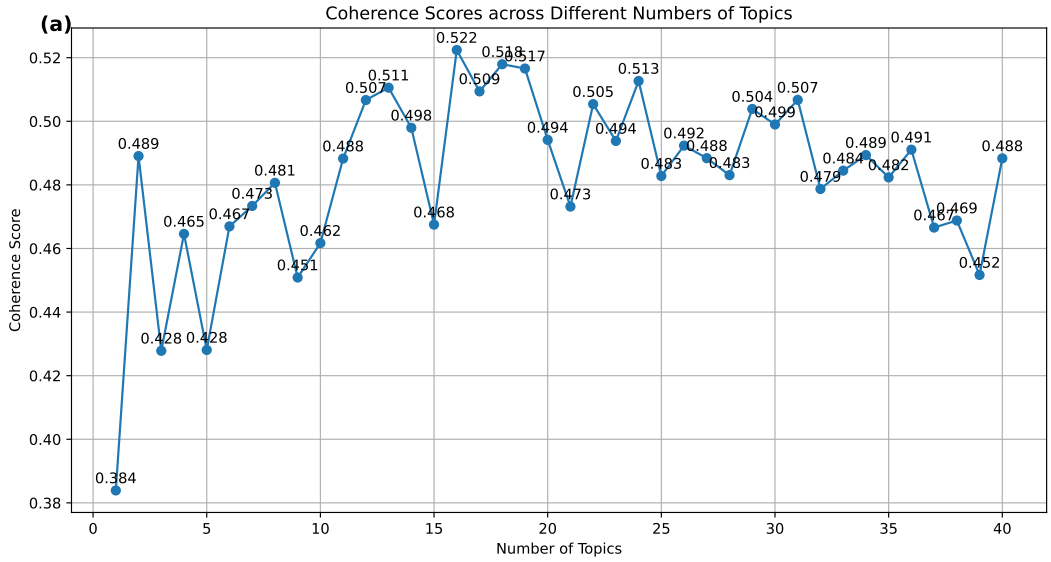

Figure1:(a)Variation of LDA Model Coherence Scores with Topic Quantity.
(b)Airbnb Listing Topic Analysis: LDA Modeling and Keyword Visualization

The LDA process will cost about 30 minutes, so wesave and re-read the output remotely from Github. Then, results shows that among 16 topics *Figure1b*, there are some topics mainly describe the location like *topic8* and *topic6*. Also, some contains information about the facilities and some adjectives towards surrounding environments like *topic13* and *topic14*. In short, all of those key words could illustrate the general features about Airbnb listings which is essential to the recommendation algorithms in platform's branding(Mody and Gomez, 2018).

6.2.2 Do the listings in the same neighbourhood, or with the same spatial location, share the similar amenities?

Amenities are highly categorizable, like '500Mb WiFi' and 'highspeed Internet access' basically meaning the same thing. Thus, we need to identify various amenities' similarities just like group synonyms out from dictionaries. We use the Word2Vec model to classify voluminous t words and phrases, and then apply UMAP(Stalder *et al.*, 2023) for better visualization in *Figure2a*.

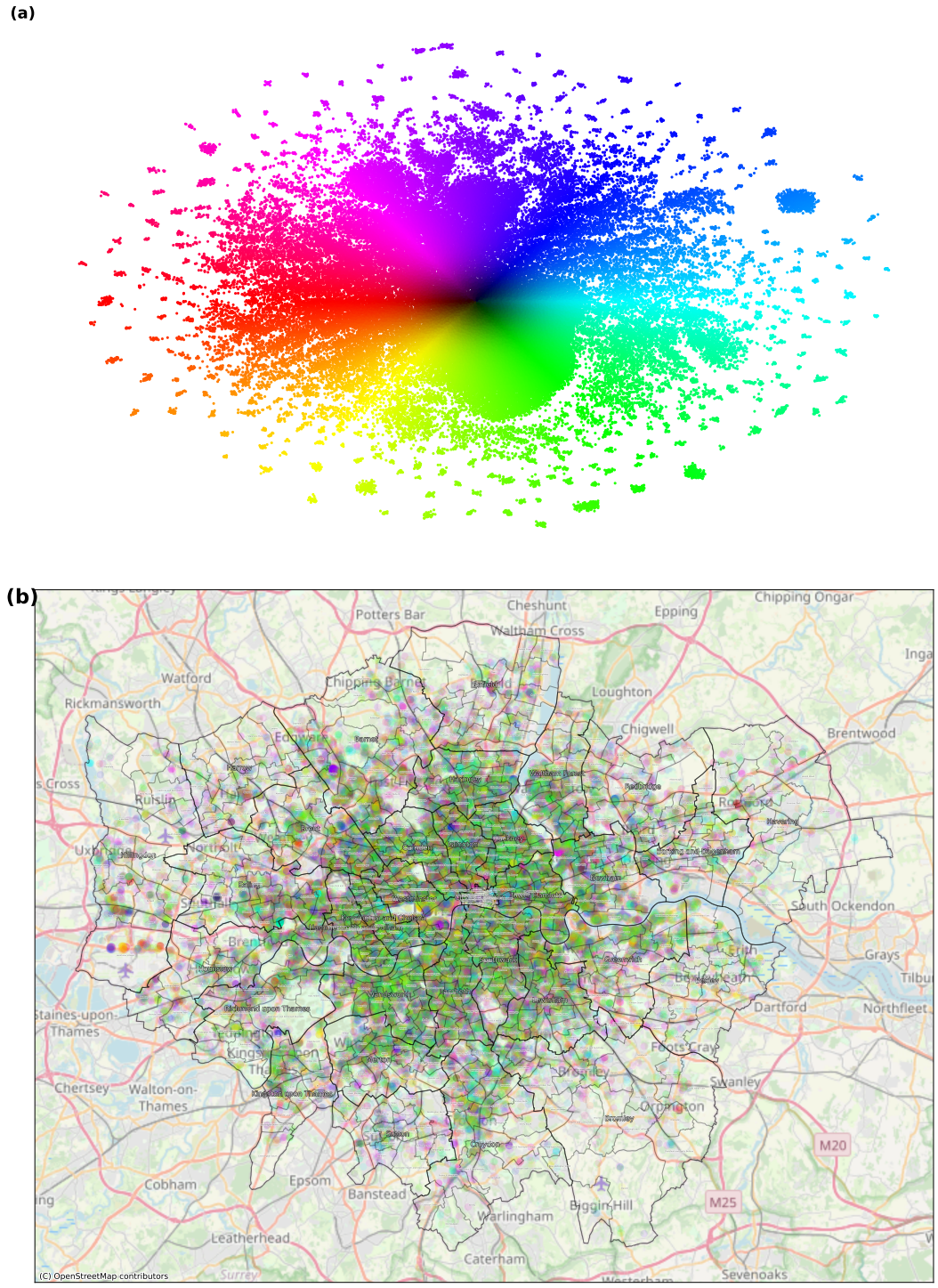

Figure2:(a)Spectrum of Features: A UMAP Clustering of Word Embeddings.
(b) Geographic Distribution of Residential Similarities in London

In the *Figure2b*, each kind of colours represents the amenities feature of a property, and areas with similar colors indicate highly similar amenities features between properties. This allows us to determine whether the properties in a specific area or community exhibit homogeneity (highly similar colors) or heterogeneity (more varied colors) in listing features.

6.3 Which indicator guide the branding?

Even though Airbnb, as a responsible company, should take community and regulation into consideration, the essence of branding and recommendation system is still aiming for profit. Therefore comes the question: what indicator could represent the potential economic opportunities for listing's branding or promotion?

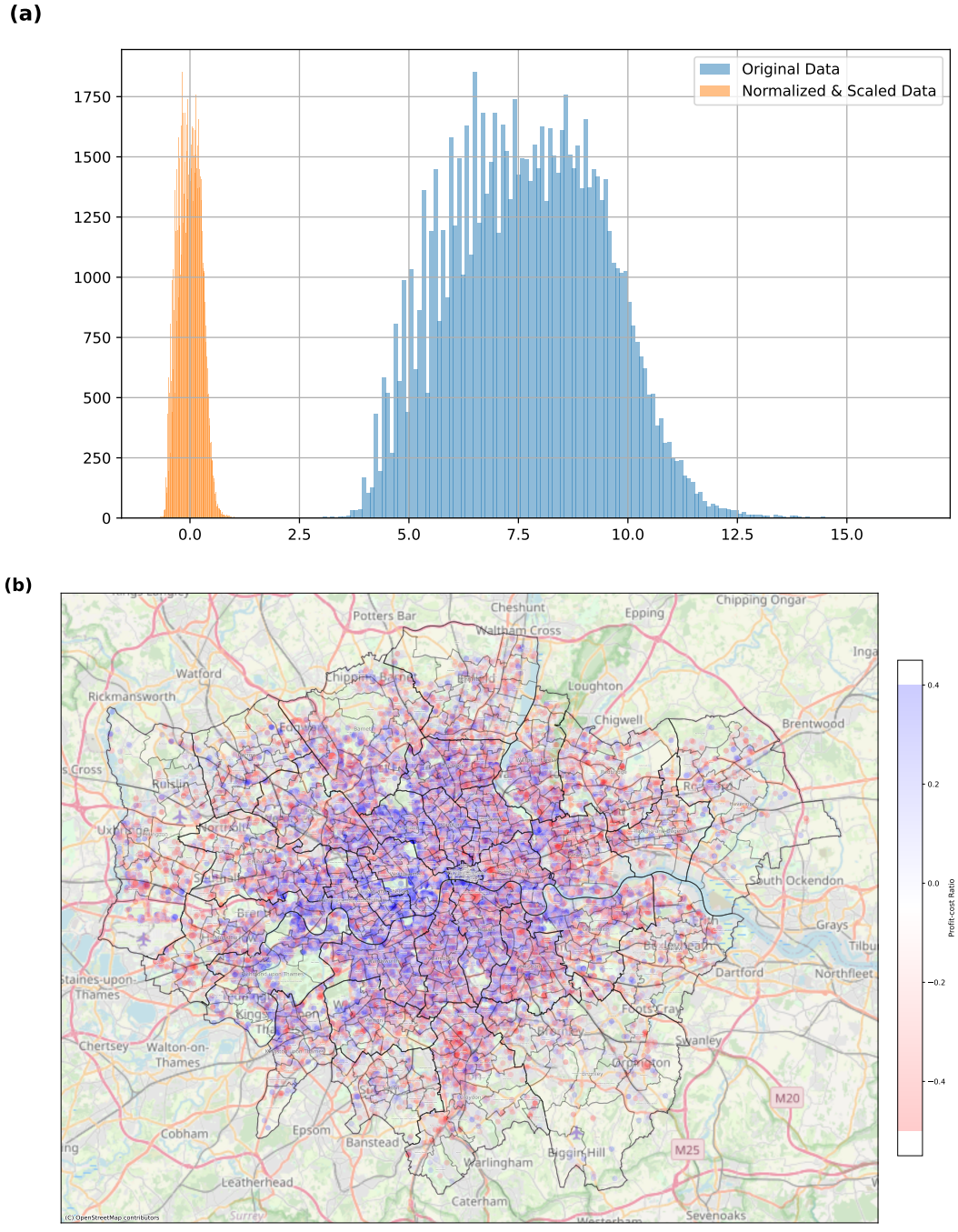

Figure3:(a)Statistical Distribution of Annual Revenue for Listings in London.
(b)Geographical Distribution of Cost-Benefit Ratio for Listings in London

We use several numeric columns to calculate the total income for every listing. Though, technically this is an approximate number with normal distribution *Figure3a*, but it aligns with the data from the Inside Airbnb. Afterwards, we compare 'sum_income' with the average in that wards to indicate this listing's 'profit-cost ratio'. Then we standarlize the data and visualize it in the map *Figure3b*. The blue

area means potential for more profit and more lease, which should be highlighted and coordinated with *Figure2b* when branding and promoting.

6.4 How does the indicator correlate with textual information?

By using the SVM model for better predicting the 'profit-cost ratio' according to the textual information, we gat an trained model with accuracy more than **85%**, which could help the Airbnb platform or Government to evaluate the listings before they are promoted and recommended to the potential renters.

6.5 Summary

After the analysis, we have the key topics and words for better generalization(*Figure1*), the features spatial distribution for better classification(*Figure2*) and the 'profit-cost ratio' spatial distribution for better investment(*Figure3*), all of which would be utilized to inform the strategies for Airbnb, landlords, communities and governments(*Figure4*).

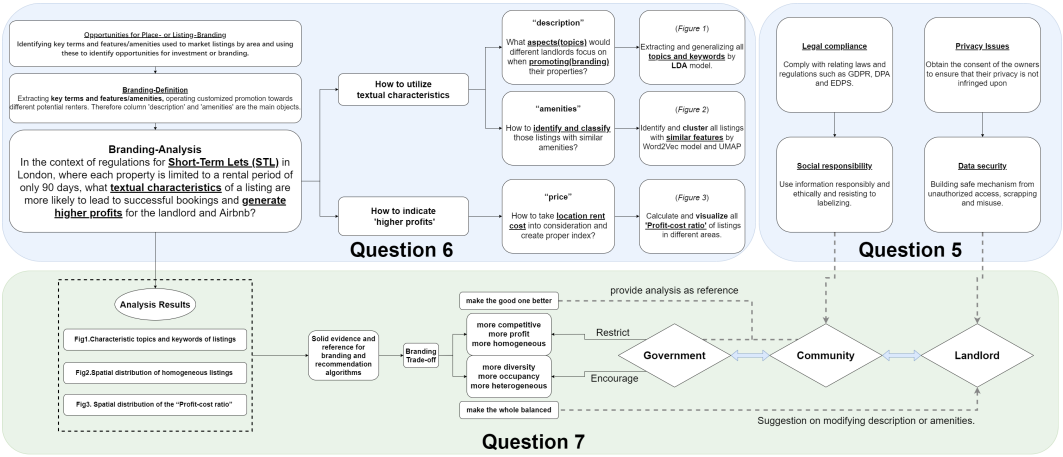

Figure 4 Framework Diagram

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

7.1 Short Term Lets(STL)

In an effort to preserve the city's current housing supply, the government legalized short-term rentals in London for a maximum of 90 days per calendar year with the introduction of the 2011 Localism Act and the 2015 Deregulation Act. Nevertheless, a number of studies(Jefferson-Jones, 2015) point out that this regulation isn't always adhered to in reality. Most of the Airbnb listings (77%) did respect the 90-day limit. Out from the listings surpassing the 90-day limit, the average estimated occupancy was 145 nights a year. Of these lettings, 6,140 (or 55%) were entire homes and 5,000 (or 45%) were private rooms. Hence, much of the existing research (Shabrina, Arcaute and Batty, 2022) has focused on the role of Airbnb as the most prominent and prevalent online platform for short-term lets in the UK and internationally.

7.2 Airbnb Branding

To enhance the Airbnb platform strategically, leveraging text features for branding and recommendation algorithms is crucial. Based on the comparison between *Figure2b* and *Figure3b* and some perspectives from Question5, The following strategies can be implemented:

7.2.1 Positive feedback cycle:

In regions with lower occupancy rates, recommendation algorithms adjustment ensures balanced occupancy rate over different areas. This proactive approach mitigates property vacancy concerns and boosts hosts' profitability, thereby fostering a dynamic equilibrium within London's housing market. Moreover, for listings with high rental profitability, providing additional positive feedback serves to incentivize competitive listings, which means positive feedback cycles, promoting business operations beneficial for both Airbnb and landlords.

7.2.2 Negative Homogeneous listing:

Considering the potential contribution of housing homogeneity to market distortions (Zhou, Gibler and Zahirovic-Herbert, 2015; Nieuwland and Melik, 2018), in areas like London Bridge & West Bermondsey, where low income rates and property feature similarity coincide, the platform and housing department should explore the incorporation of text-based features. By leveraging these features, authorities can identify and filter out homogeneous listings in concentrated areas. This strategic approach could assist platform in rationally branding homogenous properties in time series and making arrangement according to various peak demand period, as well as promoting a more balanced housing landscape.

7.2.3 Airbnb's trade-off:

In pursuing the core interests of its business, Airbnb undoubtedly seeks to foster a positive cycle by promoting competitive listings to renters (Hoffman, 2020). However, this could inadvertently contribute to homogeneity, counteracting the intended positive cycle (Hübscher and Kallert, 2022). Alternatively, Airbnb could actively contribute to the heterogeneity of specific regions by engaging in targeted interventions. One effective approach could involve providing personalized guidance to hosts in competitive areas, encouraging them to modify their amenities/descriptions to enhance their appeal to renters. As discussed in the Question 5, once Airbnb got valuable textual information, social responsibility they should take to establish a framework for communication and collaboration with hosts and provide insights towards market trends. Overall, the trade-off between promoting competitiveness and maintaining area diversity should be approached with flexibility by implementing a dynamic system that takes into account local preferences, seasonal variations, and emerging trends.

7.3 Government Regulatory Options

Furukawa & Onuki's tri-categorical definition (Furukawa and Onuki, 2019) indicates that effective policies should be less restrictive for Primary Hosted & Unhosted Short-term lets within appropriate timeframes, while regulating Non-primary short-term lets more firmly to provide the right incentives to landlords to rent long-term.

7.3.1 Tailored Policies Based on Spatial Distribution Features

Tailoring policies for diverse community types is essential. In high-density areas, consider limiting the addition of new listings to prevent overcrowding. In contrast, for areas with lower occupancy rates, policies can encourage landlords to adopt more proactive occupancy promotion strategies.

7.3.2 Dynamic Policy Adjustments for Supply-Demand Balance

Utilize spatial distribution features to monitor market dynamics and make dynamic policy adjustments based on actual demand. In high-demand areas, policies can be more flexible, encouraging short-term rentals, while in oversupplied regions, stricter policies can reduce vacancy rates. Connect the identified branding opportunities with STL regulations to strike a balance between encouraging tourism and preventing negative impacts on housing markets. For areas with distinctive amenities or features, consider regulations that preserve the uniqueness without contributing to housing shortages.

7.3.3 Encouraging Landlord Engagement in Community Development

Airbnb transforms residential communities into tourist spaces and changes the socio-cultural landscape of urban neighborhoods. It specifically propagates the experience of 'living like a local'(Ferreri and Sanyal, 2018), but this consumption of everyday local residential life has implications for the well-being of long-term tenants, including the disruption and erasure of long-term communities and housing insecurity(Rozena and Lees, 2021). Critical urbanists (Freytag and Bauder, 2018; Cocola-Gant and Gago, 2019) have accordingly linked Airbnb to touristification/gentrification - 'Airbnbification'(Törnberg, 2022). Governments can consider incentivizing landlords to participate in community development, aiming to increase the 90-day occupancy rate. This not only reduces long-term property vacancies but also fosters community vitality and helps maintain supply-demand equilibrium.

7.3.4 Create a Registration Service to Bridge Gaps in Data

In a context where the lack of data is cited as a major limitation in research and decision-making outcomes (Fonda, 2021), a registration service could provide some of the information necessary to bridge this gap. Utilizing statistical analysis and modelling, regulatory decisions can be evidence-based, considering the unique characteristics of each area. A collaborative effort between cities and Airbnb is suggested for the development of a centralized registration platform. The streamlined online monitoring and fine collection system could significantly enhance planning

authorities' ability on balancing housing prices and availability, also improving community well-being.

Reference

Bemt, V. van den *et al.* (2018) 'Teaching ethics when working with geocoded data: A novel experiential learning approach', *Journal of Geography in Higher Education*, 42(2), pp. 293–310. doi: 10.1080/03098265.2018.1436534.

Cocola-Gant, A. and Gago, A. (2019) 'Airbnb, buy-to-let investment and tourism-driven displacement: A case study in lisbon', *Environment and Planning A: Economy and Space*, 53(7), pp. 1671–1688. doi: 10.1177/0308518X19869012.

Culnan, M. J. and Williams, C. C. (2009) 'How Ethics Can Enhance Organizational Privacy: Lessons from the Choicepoint and TJX Data Breaches', *MIS Quarterly*, 33(4), pp. 673–687. doi: 10.2307/20650322.

Ferreri, M. and Sanyal, R. (2018) 'Platform economies and urban planning: Airbnb and regulated deregulation in london', *Urban Studies*, 55(15), pp. 3353–3368. doi: 10.1177/0042098017751982.

Fonda, U. D. (2021) 'Short-term lets in cambridge', *Cambridge Ahead*. Available at: https://www.cambridgeahead.co.uk/media/1959/short-term-lets-in-cambridge.pdf.

Freytag, T. and Bauder, M. (2018) 'Bottom-up touristification and urban transformations in paris', *Tourism Geographies*, 20(3), pp. 443–460. doi: 10.1080/14616688.2018.1454504.

Furukawa, N. and Onuki, M. (2019) 'The design and effects of short-term rental regulation', *Current Issues in Tourism*, 25, pp. 1–16. doi: 10.1080/13683500.2019.1638892.

Hasselbalch, G. (2019) 'Making sense of data ethics. The powers behind the data ethics debate in European policymaking', *Internet Policy Review*, 8(2). doi: 10.14763/2019.2.1401.

Hoffman (2020) 'Airbnb, short-term rentals and the future of housing', *Routledge*. Available at: https://books.google.com/books?hl=zh-CN&lr=&id=8g4HEAAAQBAJ&oi=fnd&pg=PP1&dq=airbnb+short+let+term&ots=idALQgy6ci&sig=8b38i0KyD3wMTpg1TnWJ35IwFzo.

Hübscher, M. and Kallert, T. (2022) 'Taming airbnb locally: Analysing regulations in amsterdam, berlin and london', *Tijdschrift voor Economische en Sociale Geografie*, 114(1), pp. 6–27. doi: 10.1111/tesg.12537.

Jefferson-Jones, J. (2015) 'Can Short-Term Rental Arrangements Increase Home Values?: A Case for Airbnb and Other Home Sharing Arrangements'. Rochester, NY. Available at: https://papers.ssrn.com/abstract=2714051 (Accessed: 18 December 2023).

Ji, Y., Li, H. and Yang, Z. (2021) 'An Analysis of Branding Practices of Airbnb: Implication for Future Strategic Planning', in. Atlantis Press, pp. 857–861. doi: 10.2991/assehr.k.211209.139.

La, L. *et al.* (2021) 'Location of Airbnb and hotels: The spatial distribution and relationships', *Tourism Review*, 77(1), pp. 209–224. doi: 10.1108/TR-10-2020-0476.

Mody, M. and Gomez, M. (2018) 'Airbnb and the Hotel Industry: The Past, Present, and Future of Sales, Marketing, Branding, and Revenue Management', *Boston Hospitality Review*, 6(3). Available at: https://www.bu.edu/bhr/2018/10/31/airbnb-and-the-hotel-industry-the-past-present-and-future-of-sales-marketing-branding-and-revenue-management/ (Accessed: 17 December 2023).

Nieuwland, S. and Melik, R. van (2018) 'Regulating airbnb: How cities deal with perceived negative externalities of short-term rentals', *Current Issues in Tourism*, 23(7), pp. 811–825. doi: 10.1080/13683500.2018.1504899.

Prentice, C. and Pawlicz, A. (2023) 'Addressing data quality in Airbnb research', *In-*

ternational Journal of Contemporary Hospitality Management, ahead-of-print(ahead-of-print). doi: 10.1108/IJCHM-10-2022-1207.
Rozena, S. and Lees, L. (2021) 'The everyday lived experiences of airbnbification in london', *Social & Cultural Geography*, 24(2), pp. 253–273. doi: 10.1080/14649365.2021.1939124.
Shabrina, Z., Arcaute, E. and Batty, M. (2022) 'Airbnb and its potential impact on the London housing market', *Urban Studies*, 59(1), pp. 197–221. doi: 10.1177/0042098020970865.
Stalder, S. *et al.* (2023) 'Self-supervised learning unveils change in urban housing from street-level images'. arXiv. Available at: http://arxiv.org/abs/2309.11354 (Accessed: 18 November 2023).
Törnberg, P. (2022) 'Platform placemaking and the digital urban culture of airbnbification', *Urban Transformations*, 4(1). doi: 10.1186/s42854-022-00032-w.
Voltes-Dorta, A. and Sánchez-Medina, A. (2020) 'Drivers of Airbnb prices according to property/room type, season and location: A regression approach', *Journal of Hospitality and Tourism Management*, 45, pp. 266–275. doi: 10.1016/j.jhtm.2020.08.015.
Walker, K. L. and Moran, N. (2019) 'Consumer Information for Data-Driven Decision Making: Teaching Socially Responsible Use of Data', *Journal of Marketing Education*, 41(2), pp. 109–126. doi: 10.1177/0273475318813176.
Zhang, Z. *et al.* (2017) 'Key Factors Affecting the Price of Airbnb Listings: A Geographically Weighted Approach', *Sustainability*, 9(9), p. 1635. doi: 10.3390/su9091635.
Zhou, X., Gibler, K. and Zahirovic-Herbert, V. (2015) 'Asymmetric buyer information influence on price in a homogeneous housing market', *Urban Studies*, 52(5), pp. 891–905. doi: 10.1177/0042098014529464.