CASA0013_FSDS_Airbnb_living la vida code-a

Declaration of Authorship

We, Jessica Ebner-Statt, Cerys Edwards, Jin Wen Kee, Jiayi Low, and Chung-En Tsern, pledge our honour 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 December 2024

Student Numbers: 24088089, 23197499, 19017015, 21119312, 23212203 (in order)

Brief Group Reflection

What Went Well	What Was Challenging
Pooling/Sharing information	Identifying a meaningful research direction
Delegating tasks based on strengths, interests, and experience	Identifying an appropriate metric of Airbnb occupancy
Self-directed learning for code and analysis methods	Coding for the outputs we had in mind

Priorities for Feedback

Are there any areas on which you would appreciate more detailed feedback if we're able to offer it? * Efficacy and clarity of visualisations * Methods used

Response to Questions

```
/Users/tsernian/Documents/CASA/CASA0013_FSDS/myenv/lib/python3.9/site-packages/spaghetti/network.p
  warnings.warn(dep_msg, FutureWarning, stacklevel=1)
```

1. Who collected the InsideAirbnb data?

Prior to 2015, the InsideAirbnb (IA) data (going back to 2013) was collected by Tom Slee. From early 2015, the IA data was (and continues to be) collected by founder Murray Cox, an Australian community and data activist, together with a team of collaborators and advisors comprising artists, activists, researchers, and data scientists ('Inside airbnb', n.d.).

2. Why did they collect the InsideAirbnb data?

IA data seeks to challenge official data from Airbnb, which may be misrepresentative of its operations and impact (Slee and Cox, 2016). It offers an alternative perspective to Airbnb's (limited) publicly available data by purposefully representing it through datasets and visualisations, with the not-for-profit goal of helping cities and communities to make informed decisions concerning Airbnb's operations ('Inside airbnb', n.d.). In doing so, IA increases data accessibility on Airbnb's impacts on residential neighbourhoods worldwide, especially with regard to quantifying the ramifications of short-term lets (Wang *et al.*, 2024) on local communities.

3. How did they collect it?

The IA data is collected through a process known as web-scraping, in which automated software repeatedly visits the Airbnb website and extracts publicly-available data from each listing, such as description, location, and room or property type (Prentice and Pawlicz, 2023). The Python code used to scrape the data is available to the public on Github but has not been updated since 2019 (Alsudais, 2021), meaning it is not possible to know exactly how the data are processed. However, IA does not merely scrape website data, but also processes these and augments them with assumptions about their nature ('Inside airbnb', n.d.). These approaches will be discussed further below.

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?

As a scrape of Airbnb's website rather than the raw data themselves, the final IA datasets have potential biases and quality issues that should be taken into account by analysts and legislators using them to inform policy. Web-scraping only extracts publicly-available information on Airbnb's website at the time the script is run: this means it cannot capture deleted listings or exact listing locations, as Airbnb anonymises these for privacy reasons (Prentice and Pawlicz, 2023). In addition, Airbnb's website does not differentiate between when listings are booked or blocked by their host (Crommelin *et al.*, 2018), meaning IA has to use review counts to roughly estimate occupancy rates. However, the process of scraping and processing by IA itself also introduces uncertainty. The web scrapes' reservation query settings affect the data retrieved, meaning listings may be undercounted if they do not match the search's parameters (Prentice and Pawlicz, 2023). Furthermore, Alsudais (2021) found inaccuracies in the way IA had joined reviews and listing IDs.

Moreover, it is important to remember that Airbnb's raw data is not necessarily accurate in the first place. Some listings may be fake, duplicates, or inactive (Adamiak, 2022). Finally, the IA data cannot capture short-term letting (STL) transactions through other platforms (Prentice and Pawlicz, 2023). This raises the question of whether IA data alone can provide a holistic understanding of the STL market.

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

The use of InsideAirbnb data raises a few ethical concerns due to the collection of the data through web scraping. Using an ethics framework developed by Krotov, Johnson and Silva (2020) in their paper, the ethical concerns of web scraping Airbnb's data can be categorised into infringement of individual and organisational privacy, rights of research subjects, data quality and discrimination. These categories are very applicable and in the case of IA, researchers should always be aware of identifying possible harm to individuals, organisations and enact precautionary measures to avoid these harms.

Infringement of individual privacy and rights to research subjects are perhaps some of the most significant ethical concerns while using the IA dataset. Since web scraping involves extracting all possible data from a website before parsing and classifying them, these data may unintentionally infringe on users' privacy as all web activities of individuals can be extracted, revealed and may be a means of personal identification in the future (Zook *et al.*, 2017). The IA dataset covers users reviews with their first name, duration of stay, neighbourhood, and comments recorded. Although full names and exact locations are anonymised by Airbnb, details of user reviews may reveal more about their daily lives and can risk being re-identified with generative models (Rocher, Hendrickx and Montjoye, 2019). Even if personal privacy is not harmed, users may not have given permission to researchers for the use of their data, infringing on rights of research subjects. This requires additional steps to protect anonymity of subjects by deleting identifiable information or detaching unique keys from the dataset (Kohlmayer, Lautenschläger and Prasser, 2019).

Just like how individual privacy is an ethical concern, organisations have a right to their privacy as well. Airbnb's privacy may be compromised through web scraping since their listing data embedded were not meant to be revealed entirely to the public. This may lead to confidential operations of the company being leaked including market share and intended audiences which can be maliciously used by competitors. For example, Uber was accused of using web scraping to conduct surveillance on its drivers and its competitors (Rosenblatt, 2017).

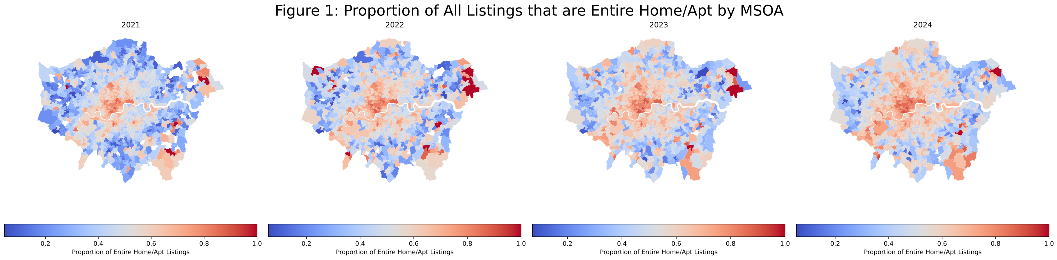
6. With reference to the InsideAirbnb data (*i.e.* using numbers, figures, maps, and descriptive statistics), what does an analysis of Hosts and the types of properties that they list suggest about the nature of Airbnb lettings in London?

Room types

An analysis of 2021-2024 Airbnb data shows a growing dominance of entire-home listings. From data on room types, we identified a rise in the proportion of entire-home listings (as opposed to single-room listings) from around 55% of total listings in 2021 to 64% in 2024.

This reflects growing demand for entire-home rentals, challenging Airbnb's claims to a "sharing economy" (**minton-23?**). An exploration of where this change is occurring (Figure 1) reveals that entire-home listings remain concentrated in central London but have steadily expanded outward over time.

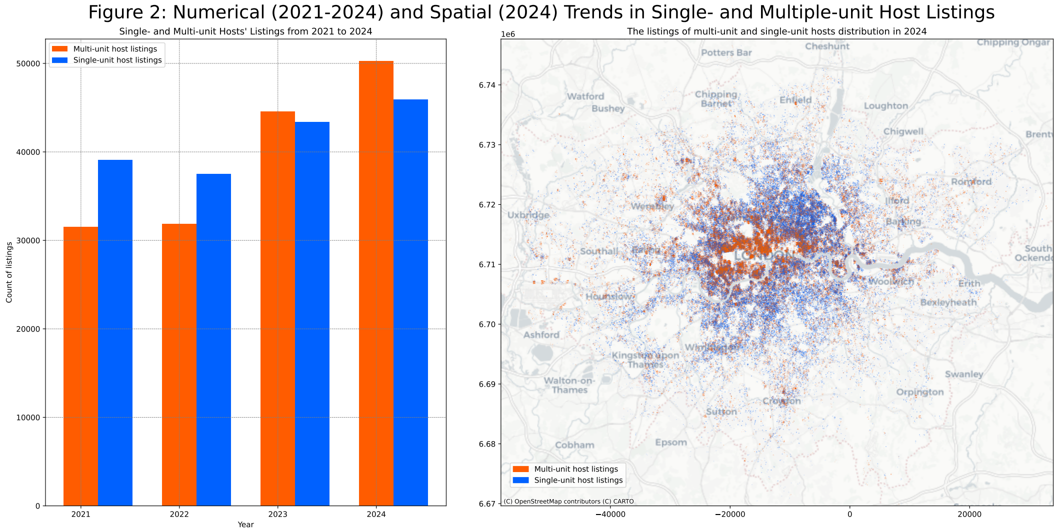
Figure 1: Proportion of All Listings that are Entire Home/Apt by MSOA



Multiple-listing hosts

Equally noteworthy is an analysis of multiple-unit hosts. As IA notes, multiple-unit hosts are likely commercial hosts (InsideAirbnb, no date), who often escape housing/land-use policies and taxation applicable to traditional landlords (Wachsmuth and Weisler, 2018), thus warranting greater scrutiny. An analysis of listings reveals that the proportion of multiple-unit host listings increased from 44.6% of total Airbnb listings in 2021 to 52.2% in 2024, reflecting an expanding dominance of the listings market. The bar chart below visualises the steady growth in the presence of multiple-unit hosts' listings; a spatial visualisation of where these hosts' properties are located (based on 2024 data) indicates a concentration of multiple-unit host listings in central London, which will be further explored below.

Figure 2: Numerical (2021-2024) and Spatial (2024) Trends in Single- and Multiple-unit Host Listings



These trends in room and host types point towards the increasing commercialisation of Airbnb lets. More than bona fide home sharing, Airbnb appears to be a platform for commercial profit at the expense of local communities (Quattrone *et al.*, 2016).

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?

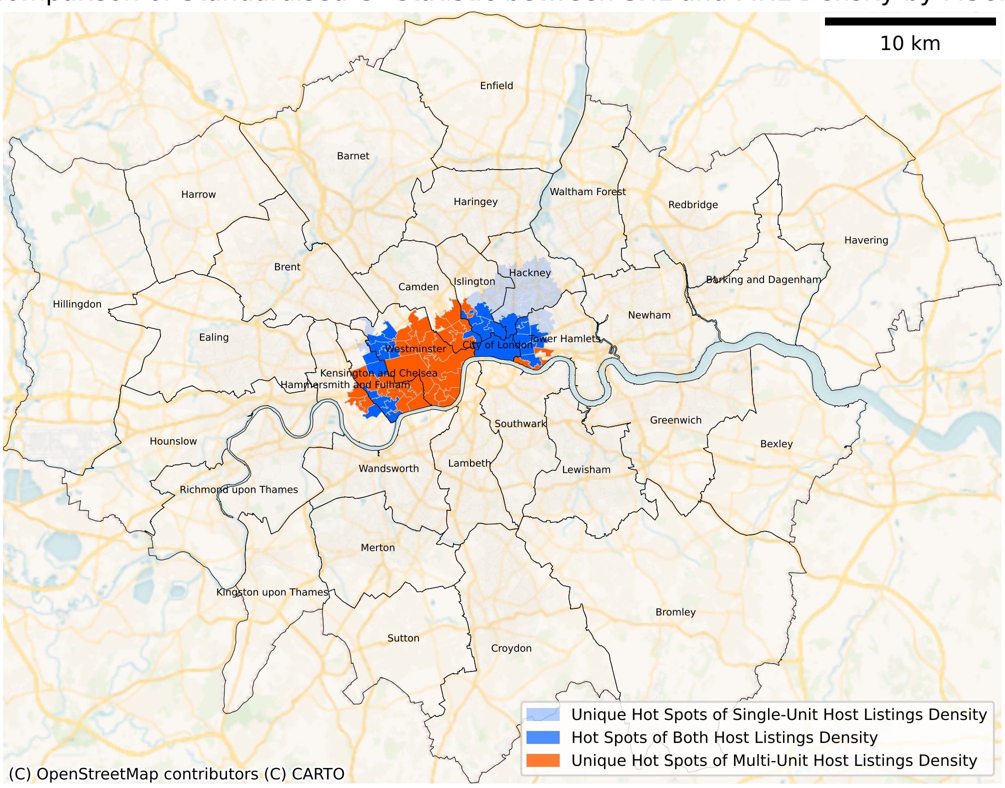
——————- Jess ——————-

Having assessed the efficacy of the 90-day policy along a temporal scale, we now consider the spatial implications of policies that aim to curb the expansion of commercialised Airbnbs in London. The effects of multi-unit hosts are well documented in the literature; Wachsmuth and Weisler (2018) reported that gentrification and reduced rental opportunities were rampant in neighbourhoods with a strong presence of multi-unit hosts, consequently transforming Airbnb's peer-to-peer sharing economy platform to a professional hosts-to-peer business operator. To identify spatial clusters of single-unit host and multi-unit host listings, we first aggregated counts of Single-unit Host Listings (SHL) and Multi-unit Host Listings (MHL) into count per square kilometers (listing density) in each Middle Super Output Area (MSOA). Getis-Ord G* statistic was then employed to determine hot and cold spots of SHL and MHL density. Due to the granularity of the MSOA layer, the G* statistic was calculated using k=8 neighbours to ensure that sufficient local patterns are captured. The G* z-scores are also standardised, making the both SHL and MHL densities comparable.

```
/Users/tsernian/Documents/CASA/CASA0013_FSDS/myenv/lib/python3.9/site-packages/esda/getisord.py:61
  warnings.warn(
```

```
/Users/tsernian/Documents/CASA/CASA0013_FSDS/myenv/lib/python3.9/site-packages/esda/getisord.py:61
  warnings.warn(
```

Comparison of Standardised G* statistic between SHL and MHL Density by MSOA



From figure 4, we observe that while hotspots of both SHL and MHL are clustered in Central London, unique hotspots of MHL are found nearer to popular tourist attractions in boroughs such as Westminster and Kensington & Chelsea. Unique hotspots of SHL are found further away in Boroughs of Islington and Hackney. This suggests that there exists a distinction of central London locations where single- and multi-unit hosts operate, with multi-unit hosts being able to operate in the most exclusive areas of central London. However, the clusters of both SHL and MHL also suggests that a spatial regulation of Airbnbs through zoning or spatial bans as seen in other cities may not be effective in curbing commercialisation of Airbnb as both clusters are close geographical proximities, with a lack of borough-level dominance of a particular listing type. This makes it difficult to single out MHL which may unintentionally disadvantage single-unit hosts who rely on Airbnb for supplemental income instead of multi-unit hosts who often operate as businesses. Instead, policy interventions could target hosts based on whether they are single- or multi-unit hosts. Multi-unit hosts could be subjected to stricter regulations, such as specific multi-unit licenses and higher tax rates, with the overall objective of assuaging the impacts of over-commercialisation of short-term lets.

Sustainable Authorship Tools

Using the Terminal in Docker, you compile the Quarto report using quarto render <group_submission_file>.qmd.

Your QMD file should automatically download your BibTeX and CLS files and any other required files. If this is done right after library loading then the entire report should output successfully.

Written in Markdown and generated from Quarto. Fonts used: Spectral (mainfont), Roboto (sansfont) and JetBrains Mono (monofont).

References

Adamiak, C. (2022) 'Current state and development of Airbnb accommodation offer in 167 countries', *Current Issues in Tourism*, 25(19), pp. 3131–3149. doi: 10.1080/13683500.2019.1696758.

Alsudais, A. (2021) 'Incorrect data in the widely used Inside Airbnb dataset', *Decision Support Systems*, 141, p. 113453. doi: 10.1016/j.dss.2020.113453.

Crommelin, L. *et al.* (2018) 'Is Airbnb a Sharing Economy Superstar? Evidence from Five Global Cities', *Urban Policy and Research*, 36(4), pp. 429–444. doi: 10.1080/08111146.2018.1460722.

'Inside airbnb' (n.d.). Available at: http://insideairbnb.com.

InsideAirbnb (no date) 'London'. Available at: https://insideairbnb.com/london/ (Accessed: 16 December 2024).

Kohlmayer, F., Lautenschläger, R. and Prasser, F. (2019) 'Pseudonymization for research data collection: Is the juice worth the squeeze?', *BMC Medical Informatics and Decision Making*, 19(1), p. 178. doi: 10.1186/s12911-019-0905-x.

Krotov, V., Johnson, L. and Silva, L. (2020) ‘Tutorial: Legality and Ethics of Web Scraping’, *Communications of the Association for Information Systems*, 47(1). doi: 10.17705/1CAIS.04724.

Prentice, C. and Pawlicz, A. (2023) ‘Addressing data quality in Airbnb research’, *International Journal of Contemporary Hospitality Management*, 36(3), pp. 812–832. doi: 10.1108/IJCHM-10-2022-1207.

Quattrone, G. *et al*. (2016) ‘Who Benefits from the ”Sharing” Economy of Airbnb?’, in *Proceedings of the 25th International Conference on World Wide Web*. Montréal Québec Canada: International World Wide Web Conferences Steering Committee, pp. 1385–1394. doi: 10.1145/2872427.2874815.

Rocher, L., Hendrickx, J. M. and Montjoye, Y.-A. de (2019) ‘Estimating the success of re-identifications in incomplete datasets using generative models’, *Nature Communications*, 10(1), p. 3069. doi: 10.1038/s41467-019-10933-3.

Rosenblatt, J. (2017) ‘Uber Data-Scraping, Surveillance Detailed by Ex-Manager’, *Bloomberg.com*. Available at: https://www.bloomberg.com/news/articles/2017-12-15/uber-data-scraping-surveillance-detailed-in-ex-manager-s-letter (Accessed: 5 December 2024).

Slee, T. and Cox, M. (2016) ‘How Airbnb’s Data hid the Facts in New York City’, *InsideAirbnb*. Available at: https://insideairbnb.com/research/how-airbnb-hid-the-facts-in-nyc/ (Accessed: 5 December 2024).

Wachsmuth, D. and Weisler, A. (2018) ‘Airbnb and the rent gap: Gentrification through the sharing economy’, *Environment and Planning A: Economy and Space*, 50(6), pp. 1147–1170. doi: 10.1177/0308518X18778038.

Wang, Y. *et al*. (2024) ‘The challenges of measuring the short-term rental market: An analysis of open data on Airbnb activity’, *Housing Studies*, 39(9), pp. 2260–2279. doi: 10.1080/02673037.2023.2176829.

Zook, M. *et al*. (2017) ‘Ten simple rules for responsible big data research’, *PLOS Computational Biology*, 13(3), p. e1005399. doi: 10.1371/journal.pcbi.1005399.