CASA0013_FSDS_Airbnb_living la vida code-a

1. Who collected the InsideAirbnb data?

(2 points; Answer due Week 7)

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.).

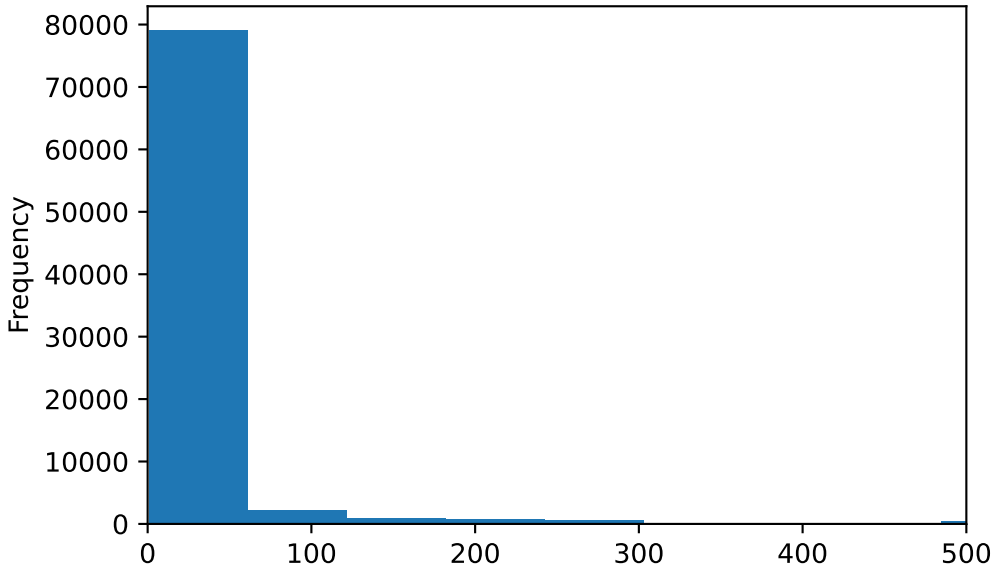
2. Why did they collect the InsideAirbnb data?

(4 points; Answer due Week 7)

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.

One of way to embed output in the text looks like this: after cleaning, we were left with 85,127 rows of data.

This way is also supposed to work ({python} f"{df.shape[0]:,}") but I've found it less reliable.



3. How did they collect it?

(5 points; Answer due Week 8)

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?

(11 points; Answer due Week 9)

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?

(18 points; Answer due **?var:assess.group-date**)

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).

Airbnb's privacy may also 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, intended audiences and other trade secrets 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?

(15 points; Answer due **?var:assess.group-date**)

Room types Firstly, analysing the categories of Airbnb listing types reveals that the nature of room types available has changed over time. Looking at room type data from 2021 to 2024, we identified a rise in the proportion of listings that were entire homes (as opposed to single-room listings) from around 55% of total listings in 2021 to 64% in 2024. This suggests an overall shift in Airbnb activity: that there is increasing demand from tenants to rent entire-home Airbnbs, and that more hosts are listing entire homes to meet that demand, deviating from Airbnb's claims to a "sharing economy" (Minton, 2023). Subsequently, an

exploration of where in London this change is occurring most significantly reveals that entire-home listings remain predominant in central London but have steadily expanded outwards over the years.

< Figure: Map showing density of entire-home listings MSOA , 2021 vs 2022 vs 2023 vs 2024 >

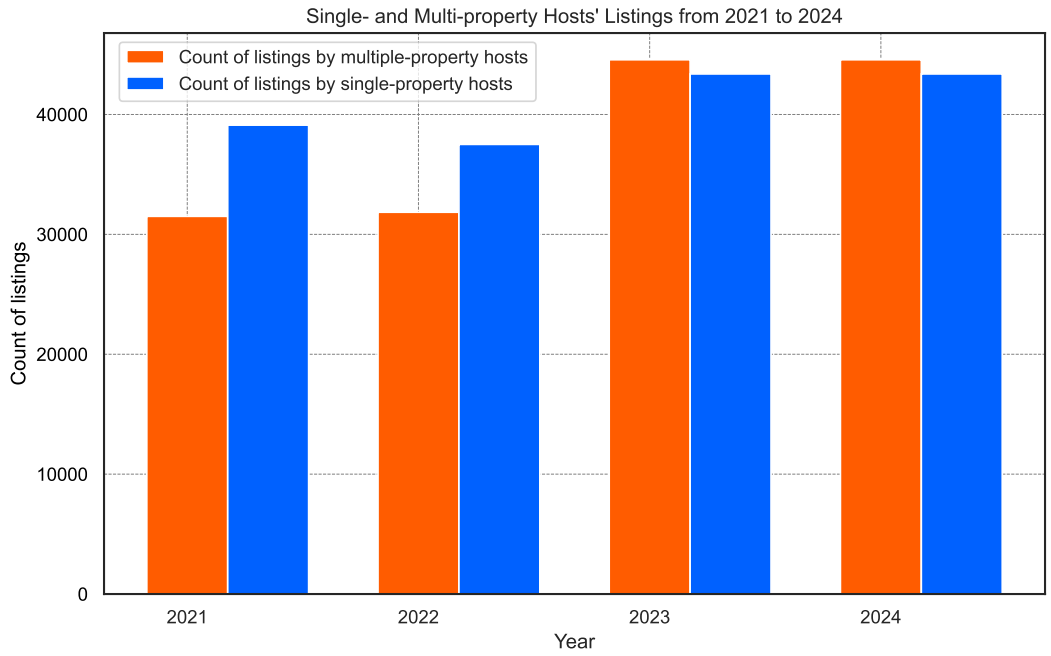
Multiple-listing hosts Equally noteworthy is an analysis of multiple-listing hosts (hosts with more than 1 room/home listed). As IA notes, multiple listings are associated with commercial hosts (Inside Airbnb, 2024), who often escape housing and land-use policies and taxation applicable to traditional landlords (Wachsmuth & Weisler, 2018) and thus warrant greater scrutiny. Our analysis of the proportions of hosts possessing single (=1) and multiple properties (>=2) revealed that the percentage of multiple-listing hosts increased from 44.6% of total Airbnb listings in 2021 to 52.2% in 2024, indicating that the proportion of multiple-listing hosts is growing to occupy more of the listings market. The bar chart below visualising the change in number of listings owned by multiple- and single-listing hosts over the past 4 years confirms a steady growth in the presence of multiple-listing hosts.

Data frame listings_2024 is 87,946 x 75
Data frame listings_2021 is 70,617 x 74
Data frame listings_2022 is 69,351 x 75
Data frame listings_2023 is 87,946 x 75
Data frame msoa_map is 983 x 13

Combined data frame is 315,860 x 76
Combined data frame is 315,860 x 15
Number of unique id is 128,095
Number of unique host id is 72,165

	host_id	year	host_listings_count	category
0	2010	2021	1	Single Property hosts
1	4775	2021	7	Multiple Properties hosts
2	4775	2022	7	Multiple Properties hosts
3	4775	2023	6	Multiple Properties hosts
4	4775	2024	6	Multiple Properties hosts
...	...	...	...	...
198747	535469107	2024	1	Single Property hosts
198748	535479813	2023	1	Single Property hosts
198749	535479813	2024	1	Single Property hosts
198750	535514014	2023	1	Single Property hosts
198751	535514014	2024	1	Single Property hosts

	id	listing_url	name	ho
0	92644	https://www.airbnb.com/rooms/92644	Rental unit in Earlsfield · ☐ 4.57 · 1 bedroom ...	49
1	93015	https://www.airbnb.com/rooms/93015	Rental unit in Hammersmith · ☐ 4.82 · 2 bedroom...	49
2	13913	https://www.airbnb.com/rooms/13913	Rental unit in Islington · ☐ 4.80 · 1 bedroom ·...	54
3	15400	https://www.airbnb.com/rooms/15400	Rental unit in London · ☐ 4.80 · 1 bedroom · 1 ...	60
4	93734	https://www.airbnb.com/rooms/93734	Condo in London · ☐ 4.62 · 1 bedroom · 1 bed · ...	49



year
2021 0.446139
2022 0.459200
2023 0.506731
2024 0.506731
Name: id, dtype: float64

< Figure 2 : Single- and Multi-property Hosts' Listings from 2021 to 2024 >

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?

(45 points; Answer due **?var:assess.group-date**)

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.

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.

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).

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.