CHASE's Group Project

Declaration of Authorship

We, CHASE, 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: 16 December 2024

Student Numbers:

24223425

24097459

24081453

24217725

22218425

Brief Group Reflection

What Went Well	What Was Challenging
Collaboration	Writing within word-count.
Evenly split workload	Merging with git

1. Who collected the InsideAirbnb data?

InsideAirbnb was collected by Murray Cox, Tom Slee, and a team of collaborators working to empower communities through data ('Inside airbnb' (no date a)).

2. Why did they collect the InsideAirbnb data?

Murray Cox and Tom Slee collected the data to critically assess the impact of Airbnb on housing markets, to provide unbiased independent publicly available data, and to facilitate the improved understanding of city authorities and regulatory bodies. Motivated by an observation of increasing entire-home listings and "multi-lister" hosts, they aimed to challenge Airbnb's portrayal as a platform for casual home sharing, revealing that much of its revenue comes from commercial operators who are pushing out local residents by raising house prices (Carville (2019)).

3. How did they collect it?

The data is collected through web-scraping, using public information from Airbnb's website.

Two main stages of data collection:

- Identify listings for chosen set of coordinates.
- Collect the following information for each listing: listing type, approximate address, number of reviews and average review score, capacity, numbers of bedrooms and bathrooms, price, and coordinates.

Data is periodically scraped for each location from the Airbnb website (Adamiak *et al.* (2019), 'Inside airbnb' (no date b)).

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 data is limited as scraping can only take place using publicly available data; which is allowed in Airbnb's robots.txt file. Datasets are therefore only an approximation of the Airbnb market and might not be suitable for use by those requiring detailed understanding of Airbnb's effect on housing markets.

Using IA's data relies solely on Airbnb's data, which only provides an estimation of the short-term rental market. Listings may be booked directly with hosts to avoid Airbnb's additional charges, appearing unavailable on Airbnb but actually booked, distorting the true effect of Airbnb on the housing market (Prentice and Pawlicz (2023)).

IA programmer Tom Slee states that "no guarantees are made about the quality of data obtained using this script, statistically or about an individual page" (Slee (2024)), encouraging researchers to check validity on their own. The script was last updated in 2019 (Slee (2024)), potentially resulting in inaccurate listing counts following changes to Airbnb's data structure, reducing the useability of data to assess housing market impacts.

The IA data is therefore not 'raw', it is "verified, cleansed, analyzed and aggregated" ('Inside airbnb' (no date b)), which introduces bias. Data cleaning erases detail and perspective, causing an issues when those analysing the data lack understanding of the context in which it was produced (D'Ignazio and Klein (2020)).

Completeness and accuracy challenges raise the question of whether researchers should rely solely on data collected by one organisation, from one website, to analyse the impact of an industry. By focusing on solely data provided by IA researchers will come to biased and partial conclusions that are influenced by the views of IA's creators.

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

Terms of Service outline the contract through which users and Airbnb interact, specifying users must not scrape to access or collect data ('Terms of Service - Airbnb Help Centre' (2024)).

From the perspective of an Airbnb user, having agreed to the Terms of Service, they expect that their data is protected. However, as InsideAirbnb demonstrates, these agreements do

not guarantee that data will not be collected. Instead, Airbnb are relying on the conscience of the programmers to adhere to the Terms of Service and robot.txt files to guarantee privacy to their users.

The Terms of Service contradicts the robot.txt file- the former specifies not scraping data at all, and the latter could be considered as permission to scrape certain data by explicitly prohibiting access to other data.

The robot.txt file is non-binding, so relies on the programmer's adherence, which could cause harm to Airbnb's users and customers if not followed.

InsideAirbnb data is collected through web scraping, which violates Airbnb's Terms of Service. Although IA takes steps to anonymise the data, Airbnb users have not provided informed consent for its collection or use by third parties. This raises ethical concerns regarding the legitimacy and responsible use of IA data (Krotov, Johnson and Silva (2020)).

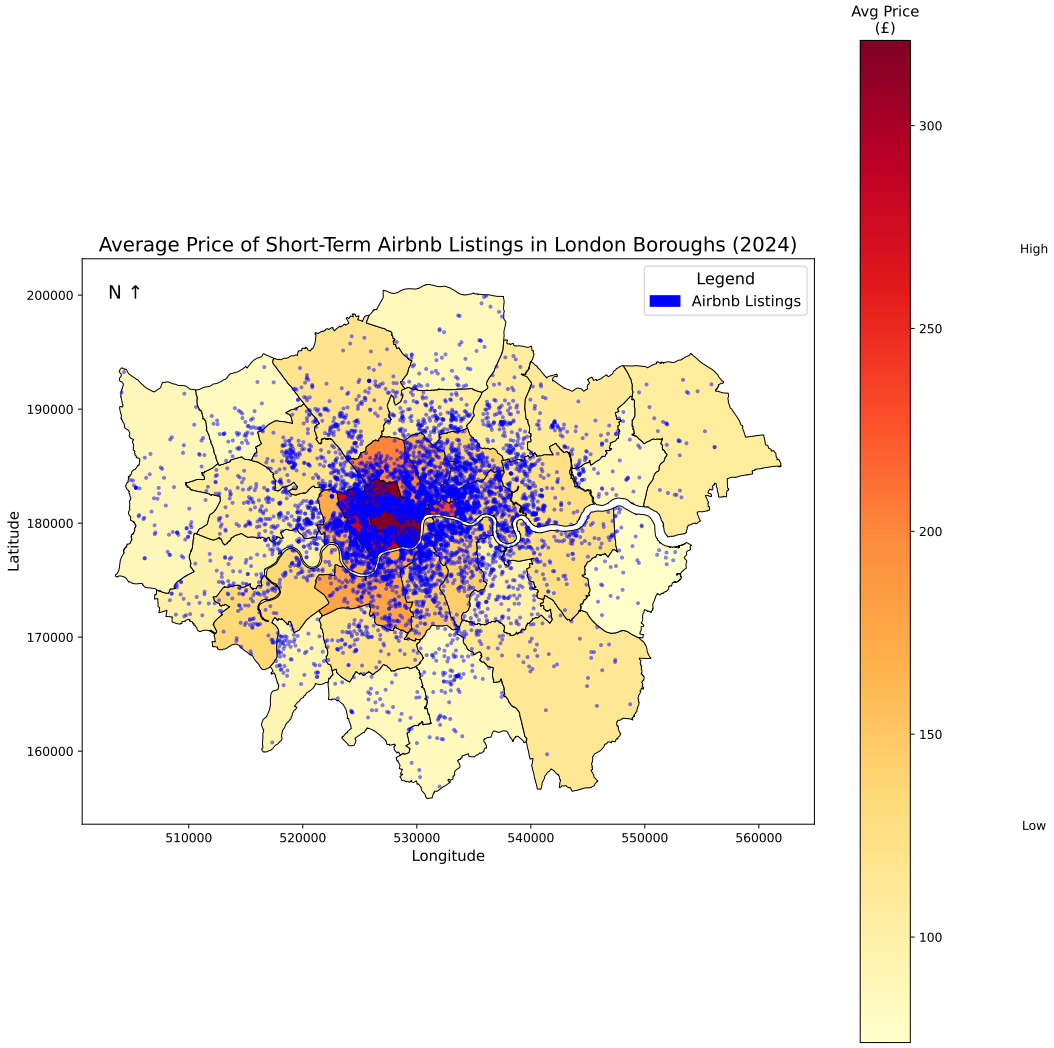
Another ethical issue is the accuracy of the data being provided. InsideAirbnb state that accuracy of the information compiled from the Airbnb site is not the responsibility of Inside Airbnb. Due care has been taken with any processing and analysis ('Inside airbnb' (no date b)).

As Mason (1986) explains, depending on how it is used, the data's accuracy can raise ethical concerns. InsideAirbnb's data is "used regularly over the last year by city analysts, journalists, academics and hospitality analysts" (Cox and Slee (2016)). Data use by city analysts will impact the lives of those that fall under the jurisdiction of the city authority, raising the ethical consideration of the data's accuracy due to the harm that may be caused from the use of Inside Airbnb's potentially misrepresentative data. This is of further concern when those analysing the data do not fully understand the process through which it has been collected and the limits this brings, causing potentially significant economic and societal impacts if InsideAirbnb's data is not used mindfully.

Mason (1986) argues that accessibility is also an ethical concern when it comes to using data. Not everyone has the technologies required to access the data, and even fewer have the intellectual skills to interpret and process the InsideAirbnb data. D'Ignazio and Klein (2020) explains that differential power has a silencing effect and quantitative data can leave people out. Whilst InsideAirbnb's mission is to "work towards a vision where communities are empowered with data and information" ('Inside airbnb' (no date b)), its capability to do so is limited if the communities it seeks to empower lack the means to make use of the available data. The lack in transfer of data science skills and knowledge to the communities that InsideAirbnb seek to represent, and subsequent reliance on external researchers, means that the imbalances in education and power will not be sufficiently addressed since the communities' reality will only ever be told through the partial perspective of said researchers.

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?

Distribution of Short-Term Listings



Airbnb's short-term listings are clustered in boroughs surrounding the City of London. 14.4% of listings occur in Westminster, followed by 8.1% in Kensington and Chelsea.

Westminster and Kensington and Chelsea exhibit the highest average prices, while Bexley and Harrow have the lowest, reflecting significant variation in short-term rental costs across the city.

Properties Available for over 90+ Nights

A concern Airbnb imposes in London is 'commercialisation'. The Greater London Authority (GLA) states that "it creates a risk of residential properties being used as letting businesses without the required planning permission and protections for neighbours" (Cromarty *et al.* (2024), p.26). To avoid this issue, homeowners are required to obtain planning permission if they intend to use residential properties for short-term accommodation exceeding 90 nights.

In the current 2024 analysis, there is a total of 6254 listings available for over 90+ nights. Westminster has 15.7% of those listings and Kensington and Chelsea at 9%.

GLA have discovered that hosts with multiple listings on Airbnb are more likely to be using the platform for commercial purposes (Cromarty *et al.* (2024), 2024, p.25). The total number of hosts with two or more listings is 6253 with the average number of listings per host being 7.06. The maximum number of listings is 1253 that belong to Sykes Holiday Cottages. This shows that the Airbnb market in London is commercialised and does not adhere to the GLA 90-night policy limit.

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?

Insights from the previous section highlight the commercialized reality of Airbnb in London despite official regulations. Failure to limit Airbnb has the potential to exacerbate the affordable housing crisis in London by reducing the availability of long-term rentals. The impact of Airbnb is of particular concern with regard to deprived households, whereby increased housing costs can contribute to displacement, making it more difficult for these residents to remain in their communities.

Our analysis aims to answer the following questions:

- What wards are "at risk" of becoming an Airbnb hotspot?
- Of these wards, which are also the most vulnerable to the negative social impacts of Airbnb?

We filtered the Airbnb dataset to include only short-term rentals (≤30 nights), recently active listings (reviews within six months), properties available ≥90 days annually, and "Entire home/apt" listings, as these have the greatest impact on housing and neighbourhood dynamics.

Decision Tree Model

We chose a decision tree methodology for this analysis due to its high interpretability, making it easy for policymakers to understand how each ward was classified as "at-risk" or "too late". Each classification decision the model makes can be easily traced in a simple, visual format.

The decision tree predicts whether a ward is likely to have high Airbnb density using ward-level characteristics of public transport accessibility, house prices, and point of interest density, which we then use to categorise each London ward into one of three groups:

Too Late: Wards already heavily impacted by Airbnb, that are in the top 5% of airbnb's per 1000 households.

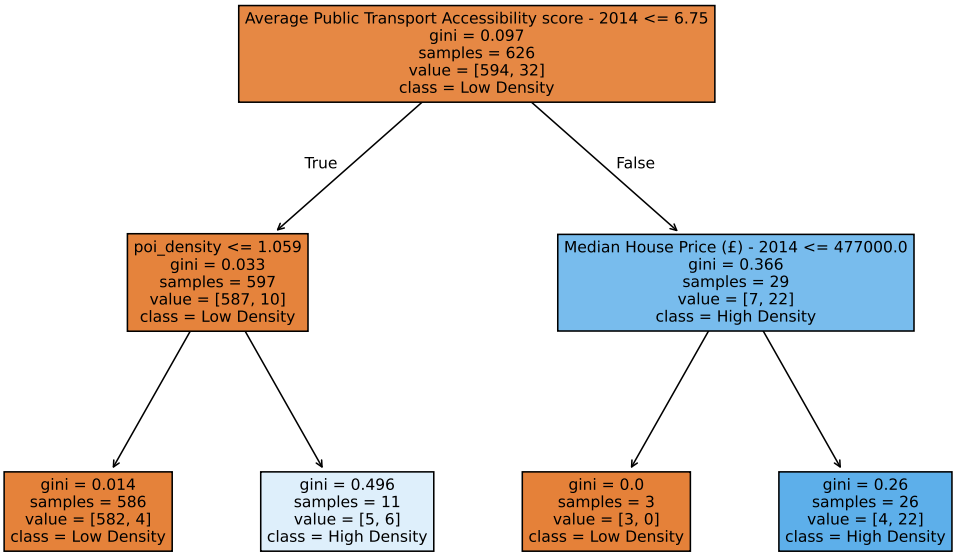
At Risk: Wards predicted to have high Airbnb density but do not meet the threshold.

Neither: Wards that don't fall into either category.

We selected these variables as research has shown that higher-income neighborhoods, better transit access, and proximity to attractions significantly influence Airbnb activity (Jiao and Bai (2020)).

```
AirBnb Density 'Too-Late' Threshold: 5.63 per 1000 households

Average Ward AirBnb Density: 1.43 per 1000 households
Average Ward House Price: £434,979
Average Ward Point of Interest Density: 0.27
Average Ward Public Transport Accessibility Score: 3.78
```



Interpreting the Decision Tree

The decision tree model identifies two scenarios where a ward is predicted to have a high Airbnb density:

- Transport Accessibility score of more than 6.75 and a Median House Price of more than £477,000

or

- Transport Accessibility score of less than or equal to 6.75 and a point of interest density of more than 1.059 per hectare.

The results suggest that high Airbnb density is linked to well-connected areas with above-average housing prices, though less connected areas can also attract Airbnb activity if they offer a high concentration of attractions and amenities.

There are 9 wards that meet either of these conditions but do not exceed the ‘too late’ Airbnb density threshold, so we categorised as ‘at-risk’.

“At-Risk” Wards and Deprivation

To assess potential social impact of Airbnb in London, we analysed the relationship between our Airbnb decision tree classifications and deprivation rank in wards. Deprivation rank is a relative measure comparing the level of deprivation across London wards. The lower the deprivation rank, the greater the deprivation score in the ward.

EDA

```
Mean Deprivation Rank for 'at-risk' wards: 200
Mean Deprivation Rank for 'too-late' wards: 319
Mean Deprivation Rank for 'neither' wards: 315
```

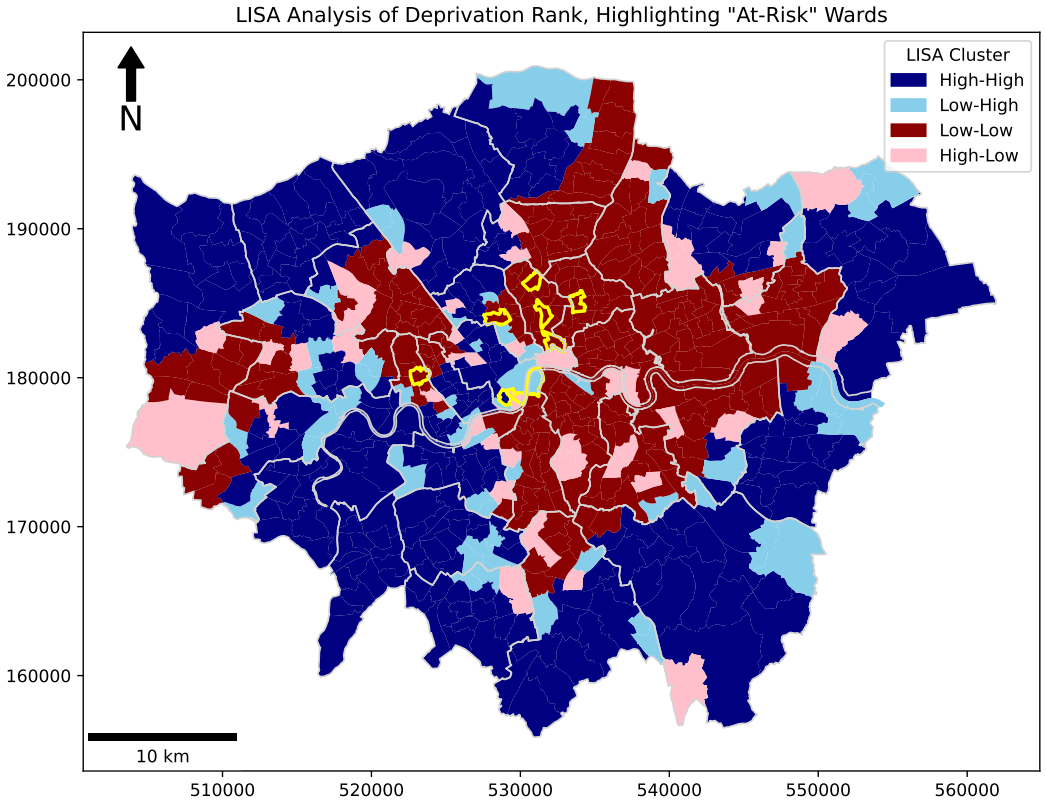
We begin with an exploratory data analysis of deprivation and ward classifications. The results shows the average deprivation rank between classification types. We see that the mean deprivation of “at-risk” wards is lower than “too-late” wards. This suggests that “at-risk” wards are not only desirable to Airbnb letters based on our classification, but are also particularly vulnerable to the negative impacts of Airbnb due to these areas being more deprived on average.

Spatial Autocorrelation Analysis

Spatial autocorrelation analysis of deprivation in London wards was employed to assess the spatial distribution of deprived wards and examine which “at-risk” wards lie within high deprivation clusters. Cluster analysis improves the generalizability of the study by identifying patterns of deprivation beyond individual wards with arbitrary boundaries.

```
Global Moran's I: 0.6888600665914413
p-value: 0.001
```

A Moran’s I test is conducted to establish that deprivation is not randomly distributed across wards in London. A Global Moran’s I statistic of 0.689 with a p-value less than 0.05 indicates there is statistically significant clustering of deprivation in London wards.



Local indicators of spatial autocorrelation (LISA) statistics allow us to visualize the clustering of deprivation in London. Outlining "at-risk" wards in yellow, we see an overlap of deprivation clustering and "at-risk" classification for three wards in Islington, one in Hackney, one in Camden, and one in Hammersmith and Fulham.

Conclusion

The study concludes there are 9 wards "at-risk" of becoming heavily saturated by Airbnb. Furthermore, six of these wards are located within clusters of high deprivation. City policy should focus on better regulating and limiting Airbnbs in the the boroughs of Islington, Hackney, and Hammersmith and Fulham to mitigate the negative impacts of Airbnb on vulnerable populations.

Limitations

Significant limitations to this study remain. We chose to use data from the 2014 London Ward Atlas in order to incorporate public transit accessibility into our analysis. However, this approach means that all other variables used from this dataset (i.e. house price and deprivation rank) are equally 10+ years outdated. Results from this study should be verified when updated accessibility scores for current wards become publicly available. Additionally, a purely quantitate analysis cannot comprehensively capture lived experience and local context. Qualitative work in our specified wards would be valuable for gaining better insights to the impact of Airbnb in these areas.

References

Adamiak, C. *et al.* (2019) 'Airbnb Offer in Spain—Spatial Analysis of the Pattern and Determinants of Its Distribution', *ISPRS International Journal of Geo-Information*, 8(3), p. 155. doi: 10.3390/ijgi8030155.

Carville, O. (2019) 'Meet Murray Cox, The Man Trying to Take Down Airbnb', *Bloomberg.com*. Available at: https://www.bloomberg.com/news/articles/2019-05-23/meet-murray-cox-airbnb-s-public-enemy-no-1-in-new-york (Accessed: 15 December 2024).

Cox, M. and Slee, T. (2016) 'How Airbnb's data hid the facts in New York City'.

Cromarty, H. *et al.* (2024) 'The growth in short-term lettings in England'. Available at: https://commonslibrary.parliament.uk/research-briefings/cbp-8395/ (Accessed: 16 December 2024).

D'Ignazio, C. and Klein, L. (2020) '5. Unicorns, Janitors, Ninjas, Wizards, and Rock Stars', *Data Feminism*. Available at: https://data-feminism.mitpress.mit.edu/pub/2wu7aft8/release/3 (Accessed: 15 December 2024).

'Inside airbnb' (no date a). Available at: https://insideairbnb.com/about/ (Accessed: 15 December 2024).

'Inside airbnb' (no date b). Available at: https://insideairbnb.com/ (Accessed: 15 December 2024).

Jiao, J. and Bai, S. (2020) 'An empirical analysis of Airbnb listings in forty American cities', *Cities*, 99, p. 102618. doi: 10.1016/j.cities.2020.102618.

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.

Mason, R. O. (1986) 'Four Ethical Issues of the Information Age', *MIS Quarterly*, 10(1), pp. 5–12. doi: 10.2307/248873.

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.

Slee, T. (2024) 'Tomslee/airbnb-data-collection'. Available at: https://github.com/tomslee/airbnb-data-collection (Accessed: 15 December 2024).

'Terms of Service - Airbnb Help Centre' (2024). Available at: https://www.airbnb.co.uk/help/article/2908 (Accessed: 15 December 2024).