Data Biography

Declaration of Authorship

I, YouTian Peng, confirm that the work presented in this assessment is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

YouTian Peng

Date: 20/11/2022

Student Number: 22076982

1 Who collected the data?

Inside Airbnb (insideairbnb.com) collected data from Airbnb around the globe-selective cities, including London. For research purposes, this independent open-source organization provides aggregated publicly available data.

2 Why did they collect it?

'Inside airbnb' (n.d.) states that data collection is for exploring and controlling the role of renting residential homes to tourists. Besides, the hosts' potential self-presentation strategies are demonstrated based on extracted data like behavior, listings, and guest reviews. From the social aspect, data collection is a method to examine conditions, enhance social cohesion and jointly plan future actions (D'Ignazio and Klein, 2020).

3 How was it collected?

Inside Airbnb directly retrieves publicly available data from Airbnb using open-source technologies, primarily Python code. The dataset's structure is constructed according to the data attributes feature. For instance, the attributes related to the review are all contiguous on the index sequence. Moreover, the dataset is controlled and tidy, following the core principles of controlling and regulating data (D'Ignazio and Klein, 2020). However, the data acquisition process did not complete the missing data.

4 What useful information does it contain?

There are a total of 74188 data, 74 attributes in this dataset. The dataset contains information about the location of all Airbnb listings in London and the host ID, hostname, room information, price, listings for each host, and availability. These data are potentially helpful for geography, economy, and sociology research.

The dataset's valuable information classifies into four dimensions:

- The accommodation's centralized or decentralized location.
- The listed unit's tangible and intangible characteristics.
- The management of the listing or unit.
- The quality of service of the host.

However, specific calendar or review data are not introduced in this dataset.

5 To what extent is the data 'complete'?

Consider data integrity from two aspects. From the perspective of quantification, 71 attributes have different degrees of missing data(see table 1), of which 21 have a missing ratio greater than 20%, and 4 have a missing ratio >99%. These four are:

- 'license'(text): the license number. Which is missing 100.000%.
- 'calendar_updated'(date): the calendar updated date. Which is missing 99.997%.
- 'bathrooms'(numeric): the number of bathrooms in the listing. Which is missing 99.997%.
- 'neighbourhood_group_cleansed'(text): the geocoded neighborhood group. Which is missing 99.997%.

Besides, among the 21 attributes whose missing percentage is greater than 20%, 15 belong to 'The quality of service of the host.' Moreover, 13 of 15 are relevant to review; the potential reason for this is that many guests do not review after check-out. Therefore, this study states that the data is partially incomplete in the quantified aspect.

As Airbnb's operations aspect, this dataset only covers some of the operation processes in London. For example, geographic data does not contain specific data describing objects and arrays or specific comments from users. These missing contents limit the operation process analysis, such as the content of review text extracted from the platform.

6 What kinds of analysis would this support?

Scholars can examine various related topics with the help of this dataset.

Firstly, this data supports the analysis of trust in sharing economy. In the sharing economy, trust reflects the accommodation service quality and is an effective mechanism to reduce the transaction cost of social exchange (McKnight and Chervany, 2001). In the Airbnb platform, the host's trustworthiness refers to reviews, rating

scores, verifications, self-descriptions, and profile photos (Zhang, Yan and Zhang, 2018). With this dataset, research about the operation of trust mechanisms and measurement of perceived trust could be proposed.

Second, the data can helps analyze the connections between the effects of the sharing economy and pricing issues. Three specific generic analysis approaches could be considered:

- To determine whether location affects pricing.
- To quantify the impact of neighborhood amenities' accessibility on pricing.
- To quantify how certain environmental factors, such as walkability and traffic noise, affect pricing.

Moreover, previous scholars studied the spatial distribution characteristics of some cities. For example, referring to a previous study by Xu *et al.* (2019), the distributions of Airbnb listings are related to the accessibility to transportation, cultural attractions, and urban centers. However, the spatial growth mechanism behind it is not well-discussed. In this case, the data support further research investigating the growth mechanism and characteristics of Airbnb's spatial distribution to deal with various related urban problems.

7 Which of the analyses outlined above are ethical?

Four ethical aspects will be discussed in this question.

Firstly, The IA data accuracy may present potential ethical concerns. While this dataset is widely used in academic research, most studies do not evaluate its effectiveness. Adamiak *et al.* (2019) captured the potential data quality problem by comparing lists collected via GitHub and IA. Meanwhile, Inside Airbnb claims no responsibility for the accuracy of the information compiled from Airbnb websites.

In addition, there are ethical issues with the transparency of Airbnb's platform data. For example, the platform disclosed its business data publicly in New York City on December 1, 2015. However, Cox and Slee (2016) suggests that the platform purged more than 1,000 "whole house" listings from its website days before publication for misleadingly describing its business. The platform's actions violate the principle of transparency and bias research results based on platform data.

Thirdly, the study of spatial distribution also involves ethical issues. Expanded geographic or social data is often required for spatial analysis. However, such data is often incomplete due to ethical reasons, for example, in San Francisco, where economic inequality has increased eviction rates since 2003. D'Ignazio and Klein (2020) noted that while the San Francisco Renters Commission collects this eviction data, it does not track where people go after being evicted or which landlords are responsible for systematic evictions in significant city neighborhoods. In this case, the lack of integrity in statistical data harms traceability in spatial distribution studies. Nevertheless, some organizations, such as the Anti-Eviction Mapping Project (AEMP), are working on mapping evictions to express dissent from San Francisco's city policy.

Finally, when examining trust or pricing issues in the sharing economy, platforms improve the user experience with less payment, still a localized experience. However, entire homes used for short-term rentals on Airbnb are taking down from the

local housing market, which could raise the rent. Moreover, some hosts gain an advantage over hotels by avoiding taxes because the hotels invest a significant part of the accommodation tax for travel promotions, which benefits all accommodation providers (Guttentag, 2015).

8 Appendix

index	num_missing	pct_missing
license	74188	100.0
neighbourhood_group_cleansed	74186	99.997
bathrooms	74186	99.997
calendar_updated	74186	99.997
host_response_rate	36409	49.077
host_response_time	36407	49.074
host_about	31694	42.721
neighbourhood	26981	36.368
neighborhood_overview	26980	36.367
host_acceptance_rate	22591	30.451
review_scores_checkin	21994	29.646
review_scores_value	21992	29.644
review_scores_location	21991	29.642
review_scores_accuracy	21955	29.594
review_scores_communication	21954	29.592
review_scores_cleanliness	21946	29.582
review_scores_rating	21903	29.524
reviews_per_month	20287	27.345
first_review	20285	27.343
last_review	20285	27.343
host_neighbourhood	17628	23.761
bedrooms	4594	6.192
description	2859	3.854
beds	989	1.333
host_location	182	0.245
bathrooms_text	159	0.214
name	21	0.028
host_listings_count	13	0.018
host_has_profile_pic	11	0.015
host_is_superhost	11	0.015
host_identity_verified	11	0.015
host_picture_url	11	0.015
host_thumbnail_url	11	0.015
host_total_listings_count	11	0.015
host_name	9	0.012
host_since	9	0.012
calculated_host_listings_count_shared_rooms	4	0.005
calendar_last_scraped	4	0.005
instant_bookable	4	0.005
calculated_host_listings_count_private_rooms	4	0.005
calculated_host_listings_count_entire_homes	4	0.005
calculated_host_listings_count	4	0.005
number_of_reviews_ltm	2	0.003
number_of_reviews	2	0.003
availability_365	2	0.003
availability_90	2	0.003
availability_60	2	0.003
availability_30	2	0.003
number_of_reviews_l30d	2	0.003
has_availability	2	0.003
maximum_nights_avg_ntm	2	0.003
accommodates	2	0.003
host_verifications	2	0.003
neighbourhood_cleansed	2	0.003
latitude	2	0.003
longitude	2	0.003
property_type	2	0.003
room_type	2	0.003
minimum_nights_avg_ntm	2	0.003
id	2	0.003
amenities	2	0.003
price	2	0.003
minimum_nights	2	0.003
maximum_nights	2	0.003
minimum_minimum_nights	2	0.003
maximum_minimum_nights	2	0.003
minimum_maximum_nights	2	0.003
maximum_maximum_nights	2	0.003
listing_url	1	0.001
last_scraped	1	0.001
scrape_id	1	0.001
host_id	0	0.0
picture_url	0	0.0
host_url	0	0.0

Figure 1: fig: Missing attribues statistical result

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.
Cox, M. and Slee, T. (2016) 'How airbnb's data hid the facts in new york city', *Inside Airbnb*.
D'Ignazio, C. and Klein, L. (2020) '5. Unicorns, janitors, ninjas, wizards, and rock stars', in *Data feminism*. PubPub.
Guttentag, D. (2015) 'Airbnb: Disruptive innovation and the rise of an informal tourism accommodation sector', *Current issues in Tourism*, 18(12), pp. 1192–1217.
'Inside airbnb' (n.d.). Available at: http://insideairbnb.com.
McKnight, D. H. and Chervany, N. L. (2001) 'What trust means in e-commerce customer relationships: An interdisciplinary conceptual typology', *International journal of electronic commerce*, 6(2), pp. 35–59.
Xu, F. *et al.* (2019) 'The influence of neighbourhood environment on airbnb: A geographically weighed regression analysis', *Tourism Geographies*.
Zhang, L., Yan, Q. and Zhang, L. (2018) 'A computational framework for understanding antecedents of guests' perceived trust towards hosts on airbnb', *Decision Support Systems*, 115, pp. 105–116.