# Who collected the data?

Inside Airbnb (IA) is a website created by Murray Cox, the website offers for direct download data that is reportedly collected from Airbnb’s website. ~~IA collects new data for each location periodically, and a new data for each location replaces all existing ones for the same location (Alsudais, 2021).~~ A variety of collaborators and partners contribute to Inside Airbnb, and it has an advisory board that assists to build sustainability for the project (*About*, no date).

# Why did they collect it?

~~IA is a website that was started by an activist who wanted to “dispute Airbnb’s claim that 87% of the hosts rent out the place in which they live” (Alsudais, 2021).~~ IA wants to provide data and advocacy about Airbnb's impact on residential communities (*About*, no date). IA wants communities to understand, decide and control the role of renting residential homes to tourists through data and information (*About*, no date).

# How was the data collected?

According to a page on IA’s website, the data is collected using python scripts, collecting public data from the Airbnb website. The collected data is verified, cleaned, analyzed and aggregated, and finally published on the IA website. IA will regularly update new data in each location (Alsudais, 2021). ~~These scripts included ones that have been “copied and pasted” from other online resources. One such resource is a script available on GitHub (Alsudais, 2021).~~

~~There are some changes that Airbnb implemented on the layout of their website and these changes negatively affected the web scraper’s performance. IA did not provide information on any possible effects these changes had on the performance of their scrapers (Alsudais, 2021).~~

# 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 existing data was incomplete or inaccurate at both the key field and record levels:

According to the disclaimer on the IA website, the location information of the listings is anonymized, which means that the accurate location of the listings cannot be obtained, which will lead to accuracy problems in exploring the geographical distribution of the listings.

If the Airbnb platform deletes the listing information, the IA website may not be updated in time, resulting in information deviation.

IA data does not differentiate between reserved and unavailable rooms, so the unavailable status of a listing may not be accurate.

There are many missing values and values with wrong data types in the records of IA data.

The data collected by IA has obvious erroneous comments and no targeted identification. And this problem may become more serious as data increases.

~~There is no research that have provided an evaluation of the IA dataset and its validity (Alsudais, 2021). However, some research suggests that data quality issues may indeed exist.~~

~~Because some websites that are not friendly to scrapers often implement methods to prevent or deceive the scripts, some of the reviews in IA data might be “spam” added by Airbnb.~~

~~The data collection code written by IA is causing the incorrect reviews problem that collect all the reviews with the specified listing ID regardless of the type of listing (“place” or “experience”). Moreover, as more experience and reviews are added to Airbnb, the quantity of incorrect data in IA is likely to increase.~~

~~Additionally, the IA dataset has many subsets and different versions, thus, a process that examines the particular subset used in a paper is required to confirm if findings reported in the paper are still valid.~~

# What ethical considerations does the use of this data raise?

Reference

*About*. (no date). Available at: http://insideairbnb.com/about/ (Accessed: 19 November 2023).

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.