1. Who collected the data? [25 words; 1 mark; *Consider the source of the data and its relation to the underlying data generating process.*]

This Inside Airbnb project data was collected by Murray Cox with the help from public volunteers and code help from Tom Slee.

1. Why did they collect it? [50 words; 2 marks; *Consider the purposes for which the data was collected and how this might shape its form/content.*]

The reason for the data collection was to help pushing the debate on Airbnb forward. This independent, non-commercial data consists of Airbnb’s business information publicly available from the Airbnb web site allow public analysis and discussion to assist revealing the real Airbnb behind the official announcement and benefit London community.

1. How was it collected? [75 words; 3 marks; *What was the method by which the data was collected and how might this shape its form/content.*]

The data set was assembled by programmatically compiling public information from Airbnb’s website. Every listing consists of consumer-facing information within London was visited, scraped, verified, cleansed and analyzed. Then the collected data was aggregated and posted as detailed listings data for London to the site named Inside Airbnb. The whole process uses multiple open source technologies: D3, [Crossfilter](http://square.github.io/crossfilter/), dc.js, Leaflet, Bootstrap, jQuery, Select2, Python, PostgreSQL and Google Fonts.

1. What useful information does it contain? [100 words; 7 marks; *A general description of key fields and records identified by the student that might support subsequent analysis.*]

This data scrapes ample information of Airbnb properties located in August London that are available from the website. Every record represents a scrape from one consumer-facing webpage of listing information display. The key fields on each listing include unique ID, listing name, description, neighborhood (overview and borough), house picture URL, host information (ID, URL, name, introduction, response time and rate, etc.), latitude, longitude, property type, room information (room type, gust capacity, bathroom number, bed and bedroom number, amenities list, price, available days) and review information (first and last review dates, review counts, scores from different aspects).

1. What useful information is it missing? [50 words; 7 marks; *A general description of key fields and records that a naive, but technically skilled, user might think that the data included but which it does not.*]

More detailed information for host, listing and review is missing. For instance, host’s identity information such as age, race, gender and nationality, host’s income from listings through a year or a month, listing’s transactions at different time scales, and review comments for every listing.

1. To what extent is the data ‘complete’? [150 words; 20 marks; *Reflecting on #4 and #5, to what extent is this data a ‘complete’ picture of the process that it claims to allow us to examine?*]

The existed data is ‘incomplete’ in both key field and record level, which remains to be identified and cleaned by users before analyzing. Some key fields are not well explained in the file. According to Inside Airbnb Disclaimers[1], the x, y coordinates and availability are not precise due to privacy protection and Airbnb calendar assuming a booked night equals to an unavailable night. The data types of fields are not all suitable for straight calculation (e.g. price), which need to be converted. As for the records, there are many rows containing missing and misplaced values.

Also, this data is ‘incomplete’ for further analysis due to missing information, like mentioned key fields in question 5 and geographical maps. Only when researchers identify and select the truly useful data, convert and clean the data, collect and add more data needed (optional) will the data be a ‘complete’ picture for further analysis.

1. What kinds of analysis would this support? [150 words; 15 marks; *Given the limitations and opportunities identified above, what kinds of analysis would this data support?*]

With the combination use of existed Boolean, numeric, text, object and spatial data in the file, a wide range of analysis would be supported after data cleaning:

The numeric data can be used for various of quantitative analysis.

Text analysis for listing description, neighborhood overview and host introduction are supported to make sentiment analysis, word clustering or word clouds.

As for spatial data analysis, combined with types of other data fields and London geographical maps, in a large city scale, the data can be used to analyze spatial patterns of listings within London to check if the distribution of points exhibit dispersed or clustered pattern. Also, in a borough scale, combined with ESRI shapefiles of neighborhood boundaries, neighborhood cleansed and other data fields, listing density, average listing price, average review number and other statistics of every borough could be calculated to see whether there is a spatial autocorrelation.

1. What kinds of analysis would it not support? [100 words; 15 marks; *Given the limitations and opportunities identified above, what kinds of analysis cannot be undertaken that a naive, but technically skilled, user might think were possible?*]

Some analyses are not supported due to data inaccuracy or data missing.

Data inaccuracy: [Numeric Data](https://github.com/jreades/i2p/blob/master/practicals/Practical-05-Numeric_Data.ipynb) analysis for availability metric will be undervalued[1]. And Small scale spatial analysis for data at the point level is not supported due to the 0-450 feet location offset[1]. For example, the number of rooms in the same building seems impossible to be counted.

Data missing: Analyses related to missing fields or data are not supported. For example, Textual data analysis for tenants’ comment and time series analysis are not supported since review text are missing and the data contains only one month’s listing scrape.

1. Which of the uses presented in Q.7 and Q.8 are ethical? [300 words; 30 marks; *Justify the ethics of these analyses with reference to examples drawn from both #7 and #8 and to the literature.*]

Ethical problems could arise in data collection and analysis[2]. As Hookway mentioned, since blogs are firmly located in the public domain, when their information is downloaded through exiting posts without interaction and intervention, the necessity of consent should be waived[3]. Generally, the data set was assembled by programmatically compiling public information from Airbnb’s blog like website, which was anonymized with no "private" information[1]. Therefore, analysis only using this data or combining this data with other ethical sources would seem to be data ethical. Below are justifications of analysis examples drawn from question 7 and 8 on whether they are ethic or not.

Textual data analyses on review comments, like cheng had explored focusing on Sydney’s Airbnb review collected from Inside Airbnb to identify users' experience[4], as well as neighborhood overview and host introduction is ethical. Because these text messages are voluntarily disclosed to the public.

However, some re-identification and classification of host’s features may trigger ‘group privacy’ problem such as age, racial, and sexual discrimination because even the protection of personally identifiable data does not protect groups[8]. Therefore, analyses requiring host’s age, race, gender or nationality information, which is private for individual and not covered in this data, are not ethical.

Small-scale spatial analysis down to the house like mentioned in question 8 will be neither accurate nor ethical since getting the precise location of the listings violates the hosts’ privacy. While large scale spatial pattern analyses such as spatial autocorrelation within London are fine. Similar researches on Airbnb’s accommodation spatial distribution had been explored to compare with hotels’ spatial distribution in Canary Islands[5] and Barcelona[6] as well as to explain and predict it’s spatial penetration in U.S. cities[7]. Part of those researches’ data source for Barcelona and U.S. are just from Inside Airbnb.

[1] C. Murray, “Disclaimers.” http://insideairbnb.com/about.html#disclaimers.

[2] M. Taddeo and L. Floridi, “What is data ethics ? Subject Areas : Author for correspondence :,” *Philos. Trans. Ser. A*, pp. 1–5, 2016.

[3] N. Hookway, “‘Entering the blogosphere’: Some strategies for using blogs in social research,” *Qual. Res.*, vol. 8, no. 1, pp. 91–113, 2008, doi: 10.1177/1468794107085298.

[4] M. Cheng and X. Jin, “What do Airbnb users care about? An analysis of online review comments,” *Int. J. Hosp. Manag.*, vol. 76, no. April 2018, pp. 58–70, 2019, doi: 10.1016/j.ijhm.2018.04.004.

[5] J. L. Eugenio-Martin, J. M. Cazorla-Artiles, and C. González-Martel, “On the determinants of Airbnb location and its spatial distribution,” *Tour. Econ.*, vol. 25, no. 8, pp. 1224–1244, 2019, doi: 10.1177/1354816618825415.

[6] J. Gutiérrez, J. C. García-Palomares, G. Romanillos, and M. H. Salas-Olmedo, “The eruption of Airbnb in tourist cities: Comparing spatial patterns of hotels and peer-to-peer accommodation in Barcelona,” *Tour. Manag.*, vol. 62, pp. 278–291, 2017, doi: 10.1016/j.tourman.2017.05.003.

[7] G. Quattrone, A. Greatorex, D. Quercia, L. Capra, and M. Musolesi, “Analyzing and predicting the spatial penetration of Airbnb in U.S. cities,” *EPJ Data Sci.*, vol. 7, no. 1, 2018, doi: 10.1140/epjds/s13688-018-0156-6.

[8] L. Floridi, “Open Data, Data Protection, and Group Privacy,” *Philos. Technol.*, vol. 27, no. 1, pp. 1–3, 2014, doi: 10.1007/s13347-014-0157-8.