DeskB's Group Project

```
'\n# Set the Github PERMALINK URL for downloading bio.bib and harvard-cite-them-right.csl\nbi
```

Declaration of Authorship

We, [DeskB], confirm 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: 19th December 2023

Student Numbers: 20017359 23032922 23081403 23103585 23130397

Brief Group Reflection

What Went Well	What Was Challenging
A	B
C	D

Priorities for Feedback

Are there any areas on which you would appreciate more detailed feedback if we're able to offer it?

Response to Questions

```
"\n# Download and read the csv file remotely from url\nhost = 'http://data.insideairbnb.com'\r
```

1. Who collected the data? (2 points; Answer due Week 7)

1. *listings.csv : This dataset was created by automatically scraping public information from Airbnb's Website. Murray Cox was one of the main founder and technicians of this mission driven project that aims to provide data and advocacy about Airbnb's impact on residential communities. [1]
2. *London_Boroughs.gpkg and London-wards-2018 : This dataset is an extract from Ordnance Survey Boundary-Line product which is a specialist 1:10 000 scale boundaries dataset.

An inline citation: As discussed on 'Inside airbnb' (n.d.), there are many...

A parenthetical citation: There are many ways to research Airbnb ('Inside airbnb', n.d.)

2. Why did they collect it? (4 points; Answer due Week 7)

:::

1.*listings.csv : Inside Airbnb is a mission driven project that provides data and advocacy about Airbnb's impact on residential communities. We work towards a vision where communities are empowered with data and information to understand, decide and control the role of renting residential homes to tourists.

2.*London_Boroughs.gpkg : With a long history and evolving from . The Ordnance Survey aims to help governments make smarter decisions that ensure our safety and security, they also show businesses how to gain a location data edge and we help everyone experience the benefits of the world outside. Under the Public Sector Geospatial Agreement (PSGA), Ordnance Survey (OS) provides Great Britain' national mapping services. OS creates, maintains and provides access to consistent, definitive and authoritative location data of Great Britain, aiming to help organisations to maximise the use, value and benefit of the data for the national interest and the public good. :::

3. How was the data collected? (5 points; Answer due Week 8)

1.*listings.csv : Inside Airbnb collects its data primarily by scraping information from the Airbnb website. This process involves the following steps:

i.Web Scraping: Inside Airbnb uses automated scripts to systematically browse and extract data from Airbnb's listings. These scripts navigate the website just like a human user would, but they do it much faster and on a larger scale.

ii.Data Extraction: Information about each listing, such as location, price, availability, number of bedrooms, reviews, and host details, is extracted and compiled.

iii.Data Aggregation: The collected data is then aggregated into a database. This database is organized to make it easier to analyze trends, patterns, and insights related to Airbnb's offerings in various cities and regions.

iv.Regular Updates: The scraping process is repeated periodically to keep the database current, capturing new listings and updates to existing ones.

v.Public Accessibility: The aggregated data is often made available to the public through the Inside Airbnb website, enabling researchers, policymakers, and the general public to analyze Airbnb's impact on housing markets and communities. It's important to note that web scraping practices, like those used by Inside Airbnb, may face legal and ethical considerations depending on the website's terms of service and regional laws regarding data privacy and usage.

2.*London_Boroughs.gpkg : "Boundary-Line for England and Wales was initially digitised from Ordnance Survey's boundary record sheets at 1:10 000 scale (or, in some cases, at larger scales). The Government Statistical Service (GSS) codes are supplied by the Office for National Statistics and General Register Office for Scotland(GROS). GIS software provides the functionality to store, manage and manipulate this digital map data. The properties of the data make it suitable as a key base for users wishing to develop applications. BoundaryLine is also suitable for use within other digital mapping systems. It's coordinated on the National Grid which allows for the easy superimposition of other data.

4. 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?

(11 points; Answer due Week 9)

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

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

6. With reference to the data (*i.e.* using numbers, figures, maps, and descriptive statistics), what does an analysis of Hosts and Listing types suggest about the nature of Airbnb lets in London?

6.1 Why should we choose the textual information?

Many studies(Xiao, 2016) have analyzed various aspects of Airbnb listings, including price, spatial distribution, room type, etc. However, the "textual description", with more impressive potential than numeric fields, also plays a crucial role in shaping renters' first impressions of the listings, contributing to facilitating successful rental transactions. Additionally, hosts would also focus on the feedback of market demands to adjust their descriptions in response to policy requirements or economic trends.

Therefore, we have the reasonable motivation to scrutinize the textual features/characteristics from the data, aiming to generalize, classify and summarize some insightful conclusion. After correlating the insights with rental potential

value, we hope to obtain valuable information about the short-term rental industry based on boroughs or wards as the basic geographical units.

6.2 What can we dig from the textual information?

Datasets consists of two textual fields: 'Description' and 'Amenities' both from the host's subjective statement as self-promotion. 'description' column is some sentences describing listing's advantages xxxxxx. 'Amenities' column is bunch of facilities and amenities inside of affiliated with the listing.

After some cleaning and preprocessing of the dataset, there are two set of questions corresponding to the two columns respectively.

1. Which topics would host like to focus on when promoting their properties?

We could use the LDA model to generalize topics and get the most frequent keywords in those topics. Firstly, we need to calculate iteratively the coherence of the LDA model with the number of topics ranging from 1 to 40, in order to determine the most appropriate number of topics for summarizing the hosts' textual descriptions (Figure 1).

```
'\n# ▯▯▯▯▯▯▯coherence▯,csv▯▯\nLDA_topic_coherence_frame = pd.read_csv("./Data/coherence_val
```

Then, word cloud shows that among 16 topics, there are xxxxxxxxxxxxx.

2. Do the listings in the same neighbourhood, or with the same spatial location, share the similar amenities?

Amenities are highly categorizable, meaning lots of similarities, like xxxxxxxxxxxx. Thus, we need to identify the similarities among various amenities from a vast vocabulary and appropriately categorize them.We use the Word2Vec model to classify a large number of words and phrases, each represented as a multi-dimensional vector. We then apply UMAP to reduce their dimensions to two (Figure 2).

In the chart, each point represents the amenities feature of a property, and points with similar colors indicate highly similar amenities features between properties. Afterward, we reposition these points on a map according to their actual geographical locations. This allows us to determine whether the properties in a specific area or community exhibit homogeneity (highly similar colors) or heterogeneity (more varied colors) in terms of amenities features.

6.3 Which indicator guide the branding?

Branding and recommendation system of Airbnb platform aims to xxxxxxx to make more money. Yet in terms of community and regulation, Airbnb should xxxxxxxx. Therefore comes the value question: what indicator could represent the potential opportunities for listing's branding or promotion?

We multiple the price of each listing by its total nights in the last year, total nights was calculated by minimum nights, maximum nights and number of reviews. Though, technically this is an approximate number, but it aligns with the data from the Inside Airbnb. Afterwards, we compare 'sum_income' with the average

property value to get an integrated index to indicate this listing's 'cost-benefit ratio'.

6.4 How does the indicator correlate with textual information?

By using the SVM to train the model for better predicting the indicator X according to the textual information, we could apply this method and get an approximately predictive result, which could help the Airbnb platform or Government to assess and evaluate the listings before they are promoted and recommended to the potential renters.

7. Drawing on your previous answers, and supporting your response with evidence (e.g. figures, maps, and statistical analysis/models), how *could* this 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

Your QMD file should automatically download your BibTeX file. We will then re-run the QMD file to generate the output successfully.

Written in Markdown and generated from Quarto. Fonts used: Spectral (mainfont), Roboto (sansfont) and JetBrains Mono (`monofont`).

References

'Inside airbnb' (n.d.). Available at: http://insideairbnb.com.
Xiao, N. (2016) *GIS algorithms: Theory and applications for geographic information science & technology*. SAGE (Research methods). doi: https://dx.doi.org/10.4135/9781473921498.