DeskB's Group Project

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?

Frankly, we've encountered lots of confusion towards the topic of this assessment.

Response to Questions

1. Who collected the data? (44words)

The dataset was collected by Murray Cox through automatic website scraping, specifically for the Inside Airbnb project. Murray Cox utilized automated tools to extract publicly available data from the Airbnb website, including location information and pricing and then compile it into the dataset.

2. Why did they collect it? (59words)

Murray Cox collected the Inside Airbnb dataset to analyze Airbnb's impact on housing markets and communities in cities globally. The project aims to provide an independent perspective, helping the public, researchers, and policymakers understand how Airbnb affects urban housing affordability and community dynamics. This data offers insights for policy discussions and social understanding of Airbnb's role in urban environments.

3. How was the data collected? (190words)

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 employs scripts to rapidly and extensively extract Airbnb listing data, mimicking human browsing. 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: Aggregated data forms a database for analyzing Airbnb trends and insights across 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: Aggregated data is often made available to the public via the Inside Airbnb website, but the web scraping it employs may face legal and ethical challenges due to website terms of service and regional data privacy and usage laws.
2. *London_Boroughs.gpkg : Boundary-Line for England and Wales was digitized from Ordnance Survey sheets at 1:10,000 scale, with GSS codes from ONS and GROS. GIS software manages this data, ideal for developing applications and compatible with other digital mapping systems. It's coordinated on the National Grid for easy data superimposition.

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? (186words)

For the listings.csv file, its data is mainly obtained by scraping information from the Airbnb website, so it may be limited by the range and depth of information publicly available on the site. For instance, detailed information about certain listings might not be fully disclosed, or website terms might restrict access to some data. Moreover, legal and ethical considerations in web scraping, such as data privacy and usage rights, may affect the integrity and accuracy of the data. The content of the

website is constantly changing dynamically, but data scraping occurs at intervals, which means the data might not be updated in real time, potentially leading to information gaps (Prentice and Pawlicz, 2023). Regarding the London_Boroughs.gpkg file, the method of data extraction relies on precise Geographic Information System (GIS) technology and detailed national geographic data. While this approach usually provides high accuracy and quality data, there may be limitations in terms of update frequency, geographic data coverage, and level of detail. Moreover, the processing and management of such data require specialized GIS technology and knowledge, which may limit the broad use and interpretation of the data.

5. What ethical considerations does the use of this data raise?(300words)

(1)Privacy issues: whether the dataset has the consent of the landlord owner to disclose its information, e.g., house location, name. Geocoded data is privacy-sensitive and highly likely to expose personal privacy when used to study demographic patterns and behaviours(**Vera_van_den_Bemt:2018?**). Therefore, obtaining the user's consent is crucial to whether the user's privacy is effectively safeguarded.

(2)Legal compliance: whether the use of the dataset complies with laws and regulations such as GDPR, DPA and EDPS.The EDPS 2015 report states that it is not enough to comply with the law in today's digital environment; we must consider the ethical dimensions of data processing.(**Hasselbach:2019?**) Therefore, legal compliance and ethical considerations should be combined in the digital age.

(3)Social responsibility: it is important to use datasets appropriately, as publishing certain data may exacerbate behaviours such as inequality and bias. The Fairness and Openness Report emphasizes how to use information responsibly and ethically, and it is quite crucial to resist the labelling of low-income communities, people of colour, etc. For example, significant differences in housing prices in different neighbourhoods may reflect economic disparities, which can affect perceptions of the socioeconomic status of those areas. Therefore, there is a need to consider how to disclose the labelled attributes of the data in question to avoid negative impacts.

(4)Data security: Some sensitive information in the dataset, such as personal information descriptions and geographic coordinates, must be stored securely to prevent access and misuse by unauthorized persons, thus avoiding security risks such as identity theft and property loss. By adjusting the norms of the use of network data, it is possible to effectively guarantee data security and increase the level of ethical behaviour of companies when processing data(**СкибунО.Ж:2022?**). Thus, attention to data security can prevent the conduct of unscrupulous individuals who collect housing data for profit or surveillance purposes.

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).

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?

STL: In an effort to preserve the city's current housing supply, the government legalized short-term rentals in London for a maximum of 90 days per calendar year with the introduction of the 2011 Localism Act and the 2015 Deregulation Act. Nevertheless, a number of studies point out that this deadline isn't always adhered to in reality. Most of the Airbnb listings (77%) did respect the 90-day limit, which nonetheless leaves an important proportion (23%) that did not. Of the listings surpassing the 90-day limit, the average estimated occupancy was 145 nights a year. Of these lettings, 6,140 (or 55%) were entire homes and 5,000 (or 45%) were private rooms.

The majority of London Airbnb hosts (84%) manage only one listing, while 1% (280 hosts) with over 10 properties each, accounting for 15% of all active short-term lettings, are predominantly commercial entities, contradicting government policy and legislation intentions. Some 72 of these companies were found to actively encourage hosts to exceed the 90-day limit with different platforms, a practice considered illegal.

Much of the existing research has focused on the role of Airbnb as the most prominent and prevalent online platform for short-term lets in the UK and internationally. Researchers, policymakers, and the public have voiced increasing concern about the possible negative externalities caused by the exponential rise in short-term lets this past decade. Local governments, in particular, are exploring viable ways to regulate and facilitate the practice while minimizing its potential negative effects.

Enhancing Airbnb Branding:

To enhance the Airbnb platform strategically, leveraging text features for branding and recommendation algorithms is crucial. The following strategies can be implemented:

- Homogeneous listing selection: Utilizing text features enables the screening of homogeneous listings, reducing the addition of new listings in spatially concentrated areas. This fosters a more balanced distribution of listings to improve overall equity in housing availability, while homogeneity might contribute to housing market distortions.
- Reinforcing recommendation algorithms in low occupancy areas: In regions with lower occupancy rates, strengthening recommendation algorithms ensures a more balanced overall occupancy rate. This proactive approach mitigates vacancy concerns, contributing to the dynamic equilibrium of London's housing market and boosting hosts' profitability.
- Positive feedback for high rental profitability listings: For listings with high rental profitability, providing additional positive feedback serves to incentivize competitive listings. This stimulates a positive feedback loop, promoting sound business operations beneficial for both Airbnb and landlords.

Government Regulatory Options:

Furukawa & Onuki's tri-categorical definition indicates that effective policies should be less restrictive for Primary Hosted & Unhosted Short-term lets within appropriate timeframes, while regulating Nonprimary short-term lets more firmly to provide the right incentives to landlords to rent long-term.

Tailored Policies Based on Spatial Distribution Features: Tailoring policies for diverse community types is essential. In high-density areas, consider limiting the addition of new listings to prevent overcrowding. In contrast, for areas with lower occupancy rates, policies can encourage landlords to adopt more proactive occupancy promotion strategies.

Dynamic Policy Adjustments for Supply-Demand Balance: Utilize spatial distribution features to monitor market dynamics and make dynamic policy adjustments based on actual demand. In high-demand areas, policies can be more flexible, encouraging short-term rentals, while in oversupplied regions, stricter policies can reduce vacancy rates. Connect the identified branding opportunities with STL regulations to strike a balance between encouraging tourism and preventing negative impacts on housing markets. For areas with distinctive amenities or features, consider regulations that preserve the uniqueness without contributing to housing shortages.

Encouraging Landlord Engagement in Community Development: Airbnb transforms residential communities into tourist spaces and changes the socio-cultural landscape of urban neighborhoods. It specifically propagates the experience of 'living like a local, but this consumption of everyday local residential life has implications for the well-being of long-term tenants, including the disruption and erasure of long-term communities and housing insecurity. Critical urbanists have accordingly linked Airbnb to touristification/gentrification - what Peters has called 'Airbnbification'. Governments can consider incentivizing landlords to participate in community development, aiming to increase the 90-day occupancy rate. This not only reduces long-term property vacancies but also fosters community vitality and helps maintain supply-demand equilibrium.

Create a Registration Service to Bridge Gaps in Data: In a context where the lack of data is cited as a major limitation in research and decision-making outcomes, a registration service could provide some of the information necessary to bridge this gap. Use statistical analysis and models to support regulatory decisions, ensuring they are evidence-based and grounded in the specific characteristics of each area. Incorporate figures, maps, and statistical insights to demonstrate the impact of STL on housing prices, availability, and community well-being.

References