Impact of Airbnb on Rental Market in London

Karol Krawiecki

October 2023

Contents

1 Introduction **2**
1.1 Context and Problem Statement 2
1.2 Research Questions and Their Relevance 3
1.3 Notation 3

2 Design and Organization **5**

3 Data Processing **8**

4 Implementation **11**

5 Querying and Conclusions **13**

Chapter 1

Introduction

1.1 Context and Problem Statement

The business model of couch surfing (sleeping over at other people for no cost) has started what we now call the 'Shared Economy'. Those models depend on platform markets and network effects in connecting people to 'share' goods or services (Ferreri and Sanyal, 2018). With couch surfing, that service was a bed to sleep on in the city we were visiting. The sharing aspect went as far as the idea that people would not charge for sleepovers, knowing that next time they would travel, they would sleep at someone else's. That idea was later expanded by the creators of the current Airbnb when they decided to make some space in their living room to set up an air mattress during the Industrial Design Conference. This short-term renting model became successful, as investors realized that it created a win-win scenario for people looking to book a short-term stay for a price less than the one of a regular hotel and provided an additional flow of cash for people who had unused space in their houses.

However, the picture nowadays is very different. With the popularity of Airbnb and the stronger focus on profitability, it appears the business could be seen as an extension of the hotel and resort industry. This comes with a set of problems, both for the cities and their citizens.

The goal of this report is to show that while Airbnb claimed to follow the original ideology of the 'sharing economy' (Törnberg, 2022), the data of Airbnb tells a different story. The pursuit of profit by the major players in the Airbnb host market brought on many challenges to the cities, among which we observe a lowering housing supply, as well as record-high rents (Wachsmuth and Weisler, 2018). To further visualize this, we will analyze the Airbnb listings in London. Given the strong professionalization trend of the hosts, fueled by Airbnb itself (Törnberg, 2022), we expect to observe strong trends in profit-oriented operations in hosts with significant listing portfolios.

1.2 Research Questions and Their Relevance

To support this claim, three questions will be reviewed in this report:

Q1. What is the average number of listings per host based on the year of joining the Airbnb platform?

Q2. What proportion of listings offer the entire property in Westminster versus Barnet for the top 3 hosts?

Q3. What is the number of listings for all neighbourhoods for the biggest host?

Question one will help specify whether there is an over-time trend of expansion of listings portfolios by the host. With the original principle of Airbnb, we would expect this average to be stable and low, between the value of one and two. This would depict the reality in which hosts engage only in letting their unused spaces, like a room or a cottage house. Inversely, given the professionalization of Airbnb, we expect to observe that the average number of listings is above 2, especially for the hosts, who are longer on the platform. This observation would help support the claim that Airbnb hosts are expanding their portfolios of listings to generate more income, rather than sharing their unused spaces.

In question two, we would like to look at the type of properties and the locations in which they are rented. Under the assumption of the original Airbnb concept, we would expect that in the central areas of London, there would be more listings for single rooms versus entire properties in the suburbs. Given the higher density of population in the city centres and smaller density in the suburbs, there should be fewer opportunities to rent an entire place in central London. Nonetheless, with the increasing trend of host professionalization, we expect to observe that there are more listings for entire properties than single rooms in central London. This issue is especially relevant for the city of London, where a huge scarcity of living spaces is observed in the central parts of London.

Lastly, I would like to focus on the portfolios of the top 10 biggest hosts in London. More specifically, I would like to understand what is the distribution of the number of listings per neighbourhood for those hosts. By understanding which neighbourhoods these hosts are targeting for their portfolio expansion, we can better observe the potential they have for creating scarcity in certain, more popular parts of London.

1.3 Notation

Whenever referring to a specific question in this report, the labels Q1, Q2 and Q3 from the list above will be used as reference. At the same time, whenever discussing *entities* and `variables` in the databases, the respective formatting shown will be applied uniformly across the whole report for better readability and cohesion. Lastly, whenever a specific SQL function will be discussed, the `blue` format will be applied.

Chapter 2

Design and Organization

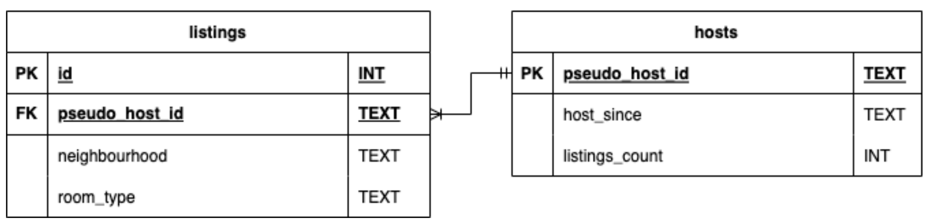

Figure 2.1: Physical ERD

Figure 2.1 represents the physical entity relationship diagram (ERD) that will be used to implement the database necessary for answering the research questions. In this figure, the data types and primary and foreign keys are declared. Attributes *id* and *listings_count* take on integer values, while *neighbourhood*, *room_type*, *host_since* and *pseudo_host_id* are considered as text.

To answer the three research questions posed in this report, one needs to look at what groups of objects are the variables referring to. In this case, the goal of the report is to try to show relationships between the hosts, the locations of their properties and the time since they joined. This specification suggests that two entities would be helpful in answering the research questions: *listings* for gathering information on the properties of the individual offers, and *hosts*, for learning about behaviours and patterns in their activity.

In principle, minimizing the number of redundant data leads to increased speed of cleaning, querying and working with a database. During the initial exploration of the data, it was also observed that more time was necessary for the authors' computer to be able to load the dataset into the used data exploration software. Instead of using a dataset with dimensions of nearly 82,000 rows and 75 columns, it is faster, as in this case, to use only 6 columns. The other variables that appeared in the original dataset did not appear here as it was deemed unnecessary for answering the research questions posed in this report.

To help answer the research questions, the 6 variables below were included in the database.

1. id
2. pseudo_host_id
3. host_since
4. neighbourhood
5. room_type
6. listings_count

Primary key `id` in *listings* allows for identifying each unique listing. Analogously, the primary key `pseudo_host_id` in *hosts* allows for the unique identification of each host on the platform. At the same time, the `pseudo_host_id` is the foreign key in *listings*, through which the reference to *hosts* is made. This foreign key set-up allows for straightforward querying of data in case we are interested in analyzing certain relations between listings and hosts (e.g.: questions this report is trying to answer).

In this case, we will be working with personal data. The attribute `host_id` from the original dataset allows us to uniquely identify each host on the platform. For that reason, pseudonymization will be necessary. An additional entity *pseudo_mapping* with original and pseudonymized host IDs will be created. The pseudonymization will be achieved by assigning random 128-bit values for each value of `host_id` from the original dataset. Using `JOIN`, a combination of pseudonymized host IDs and the corresponding data for listings and hosts will be added.

Normalization, in this case, allows for removing redundancy derived from hosts appearing multiple times for each listing they have. At the same time, normalization allows for more clarity on how is the data organized. From the original 75 variables in the single dataset **listings.csv**, we are able to create two simple entities *listings* and *hosts* of 4 and 3 variables respectively. At the same time, the simplicity of the database does not pose a significant risk to the complexity of querying necessary to connect the data, given that we only work with two tables. Moreover, understanding the 2-entities database should not be nearly as challenging as going through 75 variables presented in a simple table as was the case by using data provided directly from Inside Airbnb.

Focusing only on the variables relevant to this assignment, we observe that the dataset conforms to the first normal form (1NF). For conformity with 1NF to be achieved, attributes should only contain atomic values and have a single value (instead of a list). In this case, all 7 variables already conform to that. It is worth noting that technically it is possible to split `host_since` into a more granular form of three attributes, one for year, month and day separately. However, given the SQL functionality of `strftime`, it is easy enough to extract specific parts of those dates using basic syntax.

The second normal form (2NF) in this case is automatically achieved, as there is no composite key in this dataset (the primary key PK is a single field).

However, to achieve a third normal form (3NF), it is not allowed to have a transitive dependency of one variable on the other non-key variable. In our case, both `listings_count` (based on `calculated_host_listings_count` from original dataset) and `hosts_since` depend on the `host_id`. For that reason, a separate entity with host-specific data is necessary to ensure conformity to 3NF. The physical ERD in figure 2.1 is the representation of conformity with these three normal forms. The two tables are joined by attribute `host_id`, which is a foreign key in *listings* table and the primary key in *hosts* table. The tables have a 1:M relationship. For each listing in the table, we only have one host and each host is able to create many listings. In both cases, there is at least one host per listing and each host has at least one listing.

Chapter 3

Data Processing

Deriving valuable insights strongly relies on good data quality. For that reason, several steps will be described to provide a deep dive into what type of considerations need to be made in data management projects and the framework of the data audit that was performed in this assignment. All these steps will be performed in DB Browser.

The first check involves verifying the data type of the variables in the database. As for this paper, the software of choice was DB Browser, the important part is to verify that imported data is assigned a correct data type by the software. In this case, for each of the six variables mentioned in Chapter 1, the data types correspond to the data collected. The integer variables `id` and `listings_count` are correctly assigned their data types. Similarly, the variables `host_since, neighbourhood, pseudo_host_id` and `room_type` are correctly assigned text types.

The second part of the data inspection is related to variable names. This measure is in place to ensure that variables are not named ambiguously and follow a similar format. While the original dataset includes relatively straightforward naming, it could be still useful to rename some of the variables. For example, to differentiate between the `host_id` variable from the original dataset and the pseudonymized version, that variable was renamed to `pseudo_host_id` to indicate that those are the pseudonymized versions of host IDs. Moreover, the attributes `neighbourhood_cleansed` and `calculated_host_listings_count` were simplified into `neighbourhood` and `listings_count` respectively.

An important consideration should be made in regard to removing redundant variables from the dataset. While many columns that appeared in the original dataset are removed from the physical ERD, the original dataset is kept in a similar state as supplied by InsideAirbnb for the purpose of verifying the cleaning process (the few changes will be explained later in this chapter). In general, removing columns supports the goal of optimizing the database and increasing the speed of execution of queries.

Once the review of the table is finished, the next step is the data deduplication. It is also the first step in which we check the records of the table. From the context of how this data was collected, the `id` variable should display only unique values - each record represents a specific listing. If any record would be duplicated, that would represent a situation where a specific property is considered twice. That could potentially inflate the results of the analysis in the context of the research questions provided in Chapter 1. However, in this dataset, there are no duplicates on `id` level.

Once we ensure no duplicates, the next step is to look for missing values in the dataset. The DB browser was used to find null values in the six variables of interest for this paper. The challenge faced here was that there were five missing fields in the attribute `host_since`. Given the type of data, there were two options to tackle this issue: either remove those records since they are incomplete or to try find the missing values. Easily enough, in the dataset provided by InsideAirbnb, attribute `listings_url` (not in the general scope of this paper) redirected me to the listing page, where the year from which the host is active is available. This allowed me to fill in the missing values, which also resulted in a more appropriate representation of the hosts. In line with the format in which the `host_since` is reporting data, it was important to find out the full date of joining Airbnb. Interestingly, when trying to access the profiles of these hosts, the internet browser would bring a 404 error for every one of the hosts with a missing date of joining Airbnb as a host. Nonetheless, it is not an issue for answering Q1, as only the year of joining will be analyzed. The format used for these five instances of missing data was "YEAR-01-01". It is acknowledged that this approach might not be suitable in case the database would be used for different purposes, where month and date would have been relevant for the analysis. It is especially problematic as there would be no clear indication that those dates are wrong. Nonetheless, for the sake of time and applicability in this case, this approach is deemed suitable.

Another important aspect of data quality checks is to verify whether there are any outliers. This can hint at wrongly reported figures and could potentially impact the analysis. In this report, reviewing outliers would be especially beneficial for variables `host_since` and `calculated_host_listings_count`. For the former, it's relatively easy to identify outliers, as the year of joining Airbnb as a host is bound between the year of creation of Airbnb and the year in which the scrape was performed. For the latter variable, it would be relevant to observe the outliers to ensure that the records do not display negative figures. It would logically not make sense to have a negative number of listings on the platform. The upper bound is not as clearly defined but still should be somehow realistic.

For `host_since`, another challenge appeared in regard to how the dates are reported. In order to find all the years of the hosts in this dataset, we need to extract the year value from the string of the full date. For that reason, we had to employ the `strftime` function of SQL. Once this was done, we could easily compare the range of years for `host_since`. In this case, we saw values between 2008 and 2023, which seems plausible, as 2008 was the year in which Airbnb launched.

The next check for variable `calculated_host_listings_count` was meant to find whether we have any negative counts or if the total count was unrealistically high in any instance. In the first query, it was checked if the value of this count was 0 or negative. No table was returned, meaning there were no values fulfilling this criteria. The second query tested the top 10 hosts with the most listings to find their assigned count. In this case, the result seemed realistic.

The remaining checks are more general, as we try to find if there are any mistakes in the reported values. For example, that could involve spelling mistakes or semantic errors, where the hierarchy of objects was not followed. For the latter, it could be that the `room_type` and `property_type` were mixed (the second variable is not the focus of this study; both variables are collected in the same scraping process, which could introduce this type of error). Lastly, it is important to verify that the same type of information is written in the same way, meaning there is the same format and correct application of small/capital letters for categorical variables.

The first variable checked was `host_since`. By selecting the top five values in ascending and descending order we could inspect whether any of the values in the dataset was incorrectly inputted into the dataset. The wrong values would be moved to the front or back of the order, depending on the starting digit or the symbol (e.g.: if for some rows the date would start with day 20 - the hyphen would put it at the beginning or end of the ordered column). However, in our case, all dates were correctly formatted. The second variable `neighbourhood_cleansed` was analyzed for any typing errors or semantic violations. For the former, all the values in the domain of this variable were spelt correctly. A similar check was done for `room_type` and no issues were found. The last check was done to confirm whether the count reported in `calculated_host_listings_count` is correct. No mismatches were found, meaning that the count of listings in the dataset was done correctly. With this step done, all the checks were completed and any issues found were resolved by inputting the correct data into the dataset. In the next steps, we will be able to implement the database as described in Chapter 2 and later on, perform basic analytics to answer the research questions posed in this paper.

Chapter 4

Implementation

Before the implementation of the database as given in figure 2.1, it is important to ensure that the data management complies with GDPR. Therefore, the first step was to recognize whether we work with personal data. In this project, the original `host_id` (not a part of the final database) can be directly attributed to an individual, as Airbnb assigns a unique ID per host (InsideAirbnb, 2022). In this case, it was necessary to anonymize the data as much as possible to ensure the privacy of the hosts. At the same time, for the reproducibility of the processes employed for this assignment, it was important to allow for correlating the original hosts with the masked IDs. For that reason, I opted for pseudonymization of the data. That entails setting the database without reference to original hosts while providing a mapping table where the Airbnb IDs are paired with the universal unique ID (UUID). This way while reporting on the results of the analysis, a direct connection to the host cannot be made. It can only be achieved once the mapping table is supplied and `JOIN` function is applied with that table.

Since it was concluded that pseudonymization was necessary, firstly a mapping table was created. That table consisted of two attributes: `real_host_id` as a primary key with integer data type, and `pseudo_host_id` as text with constraints requiring values, which are unique, in each field. That mapping was then filled in with pairs of `host_id` and randomly generated 128-bit UUID values. The first challenge faced here was that SQLite was initially returning errors based on the violation of the unique constraint on `real_host_id`. The error appeared as the original *listings_cleaned* table (the table that resulted from the cleaning process described in Chapter 3) consisted of repeated values for hosts. That aligns with the initial design of the dataset as provided by InsideAirbnb and with the design of this database, as hosts can have multiple listings and thus their IDs would appear for each of their listings. For that reason, the SQL function `OR IGNORE` was applied. In each instance where a non-unique value was to be added to *pseudo_mapping*, that function allowed SQLite to ignore the entry and move on to the next host IDs. Once this function was implemented, the query worked successfully and a mapping table was created.

The next step was to create empty tables with data of *listings* and *hosts* matching the setup from the physical ERD. In Chapter 3, the data was prepared in a way that allowed me to avoid having missing values. With that in mind, a constraint was specified on all the variables of both entities that there would always be a value for each attribute. The foreign key was also specified in table *listings* that would connect with *hosts* on `pseudo_host_id`. This is also specified in figure 2.1 by the line joining the two entities.

The last step of the implementation was to fill in the tables with data. During this step, another challenge appeared, where filling in the values for `pseudo_host_id` in the entity *hosts* was violating the 'unique' constraint (which is applied automatically on primary keys). The nature and treatment of this issue were the same as described earlier during mapping table creation. Feeding the tables with data was possible thanks to SQL functions `INSERT` and `JOIN`, where the latter allows the combining of the data from multiple tables based on a set of common values. This process was applied for both *hosts* and *listings* tables and resulted in the database structure resembling the structure of physical ERD described in Chapter 2.

Chapter 5

Querying and Conclusions

I ventured into this assignment with the goal of getting a better understanding of how the major players in the Airbnb host community negatively impact the real estate market in the city of London. This fits into a bigger picture of how these hosts have confused the initial idea of the sharing economy that Airbnb is supposed to represent. Even with simple insights generated by aggregating data through SQL, I was able to clearly visualize this impact. For this purpose, the analysis provides different dimensions to the problem. The observations were made on the basis of time-trend, and location, but also a deeper look was taken into the biggest hosts in London. This is especially relevant as the biggest three hosts take up over 1,000 different properties for the purposes of short-term letting.

To this end, three questions were posed in this report. Q1 was stated as follows: "What is the average number of listings per host based on the year of joining the Airbnb platform?". To answer this question, first I focused on the *hosts* table. By applying `strftime` on `host_since` variable, I was able to extract the year the hosts joined the platform. Over those years, I plotted the average number of listings the hosts have, depending on when they joined the platform. The result of this query is shown in figure 5.1. The overall complexity of this query was low, as it did not involve the `JOIN` function due to the use of a single table, and was fully encapsulated within three lines of code. Nonetheless, the use of `strftime` and aggregate `AVG` required some additional level of knowledge of SQL. To ensure a variety in terms of depth, the first question and query were meant to be simple. At the same time, it provides a novel way of observing a trend in the development of Airbnb host community over time.

There are a few interesting insights to be made about the hosts themselves, but also about Airbnb's situation overall. Firstly, in the first two years, the were only 125 listings posted (out of which 9 were in the first year). Interestingly, it was already observed that some people were short-term renting multiple properties they had. Nonetheless, back then the impact was minimal. In the following

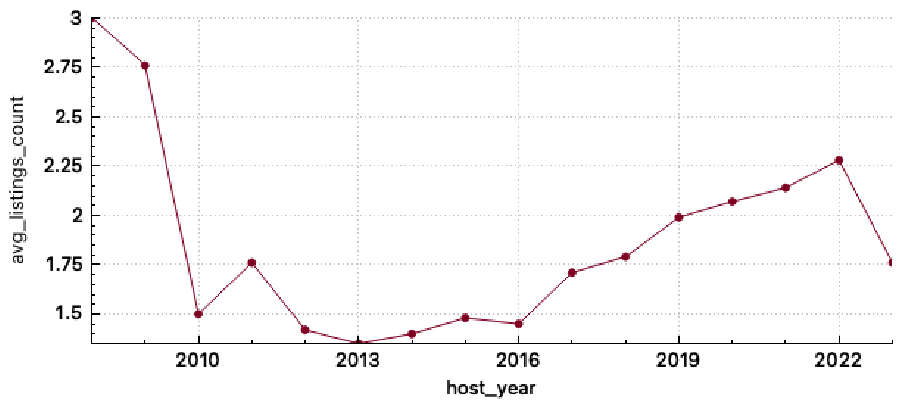

Figure 5.1: Line plot of the average number of listings of all hosts based on year of joining Airbnb

years, as more people were joining, the average went significantly down, as more people were renting individual properties or rooms. However, starting from 2016, an increasing trend in the average number of listings can be observed. It could be pinpointed to the fact that as Airbnb was becoming more popular, people started realizing the potential it has for generating additional income without the legal burden (Temperton, 2020a). Moreover, as the COVID-19 pandemic spread over the world, more cheaper listings were generated by hosts to try to recover as much as possible of pre-pandemic cashflows (Temperton, 2020b). In the end, this trend lasted until 2022 and in 2023 a decrease of 0.5 was observed. It could be argued that as the world is coming back to pre-COVID conditions, so are the Airbnb listings.

The second question in this paper was as follows: "What proportion of all listings offer the entire property in Westminster versus Barnet for the top 3 hosts?". The goal of this question was to understand what type of rental is more favoured by the hosts. Given the central location and higher population density, one might think that there will be a higher proportion of single rooms instead of entire properties. Reversely, in the outer areas of London, it would make sense that more full properties are rented out. However, in line with the professionalization trend, the expected result of this analysis is that there would be no difference in the proportion of entire properties rented out between these two areas in London. Unlike the previous query, the query involved in answering this question could be considered difficult. The reason for that is the type of insight this question aims to generate. For this query to work, aggregation on the host and neighbourhood level needs to be made. On top of that, an average number of entire properties against all room types was calculated per each combination of host and neighbourhood. Not only a `JOIN` was needed, but a subquery was used to extract the top 3 hosts in the database.

The result of the whole query can be found in the table 5.1.

Table 5.1: Proportion of 'Entire home/apt' Listings for the Top 3 Hosts in Westminster and Barnet

Pseudo Host ID	Neighbourhood	Proportion
1882F151D7937670B802F6289F659F41	Westminster	0.79
1882F151D7937670B802F6289F659F41	Barnet	0
753355E2B2577286BC96694076441E4B	Westminster	1
753355E2B2577286BC96694076441E4B	Barnet	1
E1BAD51B2EC5981772901103FCA7B8C9	Westminster	1
E1BAD51B2EC5981772901103FCA7B8C9	Barnet	1

As expected, it can be observed that entire homes or apartments make up the majority or the whole portfolio of short-term rental real estate. For the first host on the list, 79 per cent of their listings were for renting an entire property. Given that this host has over 200 listings in Westminster, this is a troubling figure. At the same time, the host did not have any properties in Barnet, which is the reason this proportion is equal to 0. For the second and third host, the situation is even worse, as all their properties in these neighbourhoods are listed for rent as entire spaces, instead of individual rooms. With over 500 listings across all of London, where almost all listings are rented as a whole, it can be observed how few individual entities can fuel the scarcity of properties in London.

Lastly, the third question was stated as follows: "What is the number of listings for all neighbourhoods for the biggest host?". The goal of this question was to look in detail into the geographical distribution of listings across the whole of London for the biggest host. Since they have over 500 listings, it can provide an interesting insight into an optimal listings portfolio for profit-driven hosts. If it is assumed that most hosts would have similar targets in mind, it could provide additional evidence into why there is such scarcity in London, especially in central areas. The query used for finding these insights could be considered medium-to-hard difficulty, partially as in principle it follows a similar pattern as Q2. The top host is selected using a subquery, however, instead of specifying the neighbourhoods of interest, we allow SQL to generate a table of all listings within the portfolio of the biggest host. The resulting information was plotted over a bar chart, as shown in figure 5.2.

The x-axis represents each of the neighbourhoods in which the biggest host has at least one listing, while the y-axis shows the number of listings in specific neighbourhoods. The clear outlier in this table appears to be the neighbourhood of Westminster. For a profit-driven host, it makes sense to provide as many properties as possible in a key location in London, where there is easy access to tourist attractions, transportation and restaurants. More surprisingly, that number is more than double as in Brent, the second biggest neighbourhood in terms of number of listings for that host. Brent, as well as Camden and Lambeth,

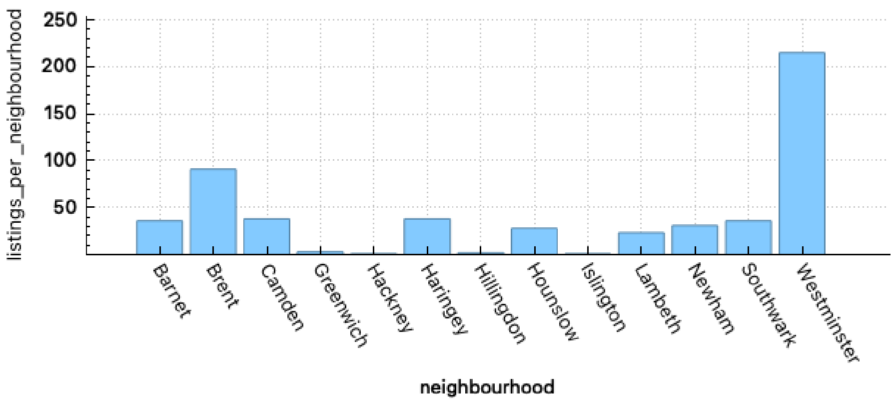

Figure 5.2: Distribution of number of listings per neighbourhood for the biggest Airbnb host in London

could be considered other central locations of London. Nonetheless, it is also visible that other, more suburban neighbourhoods of London, are a substantial part of this host's portfolio. Perhaps properties in this location offer bigger space, making it somewhat of an interesting choice for potential renters with families or groups.

While these three research questions provide a basic insight into the presence and magnitude of Airbnb in London, the key message should be that there is a reason to worry about the behaviour of Airbnb hosts and the lack of ethical consideration of the complications of the business model the platform and Airbnb as a company generates in its current state. There is, however, hope that things will take a turn for the better, as more people start to show opposition to this phenomenon, causing the local government to act against further development of profit-driven short-term rentals through Airbnb. Another positive aspect of the protests is that they caught the attention of the CEO of Airbnb. In a recent article, he openly communicated that the direction in which Airbnb is currently going is opposing the initial principles it set for itself back in the 2000s (Ekstein, 2023). The question that follows is how honest was he and what will be the effects of his improvements in London, but also in other cities globally.

This project set out to further understand and explain some of the troubles generated by Airbnb for the housing market in London, nonetheless, it is important to acknowledge the limitations of this paper. Firstly, more research is necessary to confidently confirm or reject the claims made in Chapter 5. Moreover, one needs to consider the quality of the data collected through the scraping process of InsideAirbnb (Prentice and Pawlicz, 2023). While for the sake of this assignment, it is assumed to be of sufficient quality, one cannot consider this data as the full picture.

Bibliography

Ekstein, N. (2023, October). Airbnb is fundamentally broken, its ceo says. he plans to fix it. https://www.bloomberg.com/news/articles/2023-10-02/airbnb-is-broken-its-ceo-says-here-are-his-plans-to-fix-it

Ferreri, M., & Sanyal, R. (2018). Platform economies and urban planning. *Urban studies (Edinburgh, Scotland)*, *55*(15), 3353–3368. https://doi.org/10.1177/0042098017751982

InsideAirbnb. (2022, August). Inside airbnb data dictionary. https://docs.google.com/spreadsheets/d/1iWCNJcSutYqpULSQHlNyGInUvHg2BoUGoNRIGa6Szc4/edit#gid=1322284596

Prentice, C., & Pawlicz, A. (2023). Addressing data quality in airbnb research. *International journal of contemporary hospitality management*. https://doi.org/10.1108/IJCHM-10-2022-1207

Temperton, J. (2020a, February). Airbnb has devoured london – and here's the data that proves it. https://www.wired.co.uk/article/airbnb-london-short-term-rentals

Temperton, J. (2020b, March). London's rental market is being flooded by bargain airbnb listings. https://www.wired.co.uk/article/airbnb-coronavirus-london

Törnberg, P. (2022). How sharing is the "sharing economy"? evidence from 97 airbnb markets. *PLOS ONE*, *17*(4), e0266998. https://doi.org/10.1371/journal.pone.0266998

Wachsmuth, D., & Weisler, A. (2018). Airbnb and the rent gap: Gentrification through the sharing economy. *Environment and planning. A*, *50*(6), 1147–1170. https://doi.org/10.1177/0308518X18778038