Individual Integrative Assignment

Maksimilian Štajer – Student number 682271

Table of Contents

Task 1

Problem statement and research questions

Over the years, Airbnb has transformed from a straightforward home-sharing platform into a comprehensive short-term rental platform. This transformation has been characterized by a shift in the nature of listings, transitioning from primarily lived-in homes or rooms not currently in use to properties acquired and furnished exclusively for the purpose of Airbnb listings. Moreover, it has witnessed the emergence of businesses and hosts specializing in Airbnb rentals. To maintain clarity and coherence throughout this research paper, hosts with a single listing will be denoted as "single hosts," while hosts with multiple listings will be referred to as "multi hosts."

Airbnb is part of the larger emerging wave of the “sharing economy”, focused on the sharing of homes or rooms in this case. The paper “How sharing is the “sharing economy”? Evidence from 97 Airbnb markets” found evidence that a majority of the market revenue from AirBnb tends to go to 10% of the hosts. Certain jurisdictions have passed regulations in order to limit this centralization however, these measures have so far only managed to slow the process down instead of stopping it. (Törnberg, 2022)

To see if AirBnB is still a platform offering cheap short-stay rentals providing homeowners with some additional income or has it become just another quasi-hotel offering platform enabling the increasing concentration in home ownership, and has the coronavirus changed that. I will be focusing on the following questions:

1. How has the proportion or number of multi hosts changed over the years and did the emergence of the corona virus in 2019, change the trend?
2. What is the geographical distribution of listings, categorized by whether they are managed by single hosts or multi hosts, additionally is price affected by the neighbourhood the listing is in?
3. Is there a statistically significant disparity in average ratings between listings overseen by single hosts as opposed to those managed by multi hosts?

The first question seeks to examine how the prevalence of multi-hosts on Airbnb has changed before and after the emergence of COVID-19. It could shed light on whether the pandemic affected the willingness of hosts to rent multiple properties, potentially indicating economic or social shifts related to the crisis.

The second question explores if multi hosts are geographically distributed in similar neighbourhoods, which might suggest that multi hosts focus their listing on certain more ‘touristy’ and central areas of London, making the housing crisis in those areas even worse. Looking at price and location together will help us answer if the multi hosts focus on more expensive areas and if there is significant difference in price holding location constant between single hosts and multi hosts. It must be mentioned however that price is affected by many other factors that we will not be delving into in this analysis, therefore any conclusions will have to take that into account.

The third question delves into whether there is a statistically significant difference in average ratings between listings managed by single hosts and those overseen by multi hosts. Such disparities could point to varying levels of service quality or customer satisfaction, which have social and economic implications for both hosts and guests.

Additionally, the number of listings by multi hosts can be a microcosm of the general housing crisis, with richer individuals or businesses buying multiple properties, driving prices up and forcing middle class individuals into renting instead of buying. London especially has had a problem with housing and rent prices becoming increasingly out of reach for the average person. This is related both to an increase in demand and a decrease in supply (Norwood, 2022). The decrease in supply could be related to an increase in homes owned by a small group of businesses or individuals, among which we can also find multi hosts.

Task 2

ERDs

Conceptual model

Based on the set research questions and looking at the available data, I have decided to incorporate 3 entities into my ERD, namely Host, Listing and Rating. I was also considering adding another entities, namely Location, however as Location only has one relevant column, neighbourhood_cleansed, I believe it can stay part of the Listing entity. Entities Host and Rating will share a primary key as the attributes of the Review entity we are interested in are unique for each listing. Instead of review I named my entity Rating, as I feel it more accurately conveys what the attributes of the table will contain, which I will elaborate on in the logical model chapter.

The relationships between the entities are as follows:

- Each listing can have only one and has a minimum of one host, while a host can have a minimum of one listing but can have many listings.
- Not every listing has to have a rating, while every rating has to be attributed to a single listing.

Luckily, we do not have any many to many relationships we would have to resolve.

Logical model

From the primary data I selected the attributes I felt were relevant to answering my research questions. The Listing entity contains data focusing on price and neighbourhood of the listing and the associated host for each listing.

For the Rating entity I selected all quantitative rating data as well as the number of reviews per listing and the number of reviews in the last year. The primary key of the Rating entity is the same as the primary key of the Listing entity, as it allows to connect the tables neatly while retaining the characteristics of the primary key.

The Host entity contains the number of listings each host owns. We also have data that tells us when the host first registered on the platform, allowing us to compare the ratio of multi hosts, before and after 2019.

Physical model

As we have no many to many relationships to resolve the physical model only needed us to specify the foreign keys.

In the Listing entity the only foreign key is the host_id which allows us to connect with Host entity. In the Rating entity the listing_id is both the primary key and the foreign key connecting the Listing entity with the Rating entity.

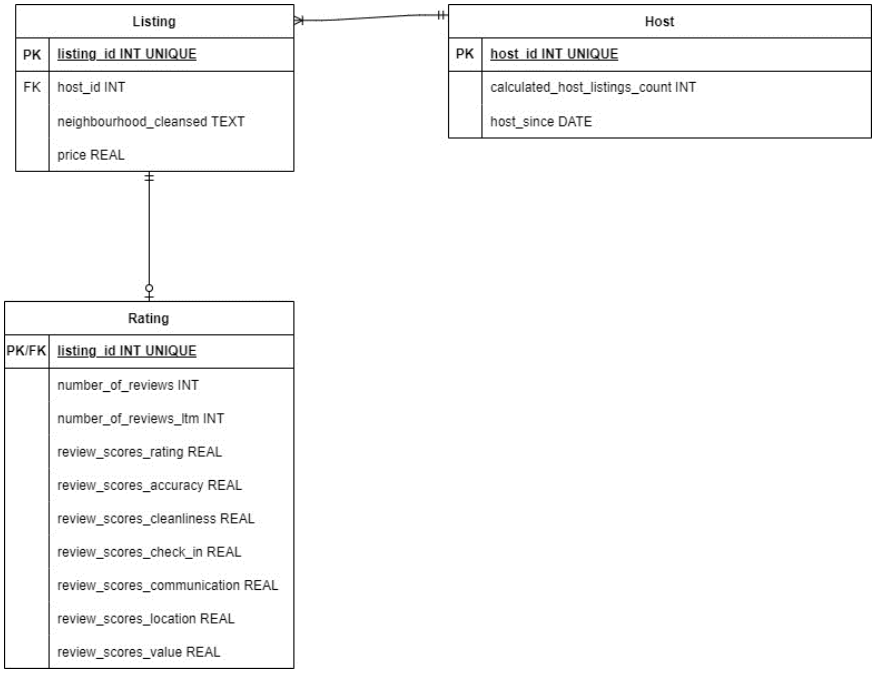

Figure 1 Physical ERD

Task 3

Cleaning the data

Cleaning the data involves checking the 6 dimensions of data quality:

- Accuracy
- Completeness
- Consistency
- Timeliness
- Traceability
- Accessibility

Accuracy, completeness and consistency are addressed in this part of the research, while traceability is addressed in task 4. Timeliness is not

The first step I did before starting with the cleaning is make a back up of the original database, so I could easily go back to the original database and run my queries again if I messed up during the cleaning process.

I checked for duplicate values in the id column and found none so I proceeded to check my chosen attributes for null values. The first attribute with null values I found was host_since, after looking at the whole rows of the 5 null values I noticed there is no host data whatsoever and also no data on the listings ever being reviewed. It seems to have been an error during the scraping process, but I decided to remove the 5 rows, as it would not have a significant impact on my sample size while ensuring the completeness of my data.

After looking at review_scores_rating I noticed more than 20000 null values, but as I decided to put those values in a separate Rating entity, I would be able to get rid of those null values, while not removing any otherwise complete rows from the Listing entity. The other more worrying null values

were the null values of the other more detailed review_scores, but looking through the data I noticed that most of the detailed review scores were the same as the overall review_score_rating, so I decided to change the null values of those attributes into the review_scores_rating of the same row.

Before moving on to pseudonymizing the personal data, I changed the price column to the REAL data type, removed the currency sign in the data and removed the comma as a thousand separator.

Pseudonymizing the data

The personal data needed to be pseudonymized was host_id and id (listing id). To pseudonymize the data I used two approaches. For the listing id I created a new table called Mapping_auto with an autoincrementing primary key and a column for the old id, then I populated it with the id from the listing table and I got my pseudonymized data.

For the host_id I encountered more problems as there are duplicate host_id present in the listings table. Due to that I first created the table Mapping_host_id with old_host_id as the primary key and new_host_id as attribute. I first populated the old_host_id column with the distinct host_id from the listings table and set a temporary value for the new_host_id, then I used the hex(randomblob(16) statement to set new ids for every host.

With that the cleaning and pseudonomyzing is complete, we used the 'VACUUM' statement to extract the cleaned database and made a back-up, now we can carry on to the creation of the entities as specified in the physical ERD.

Task 4

Processing

During task 4 I had to create and populate the tables as I envisioned them in the ERD. Before starting on that however I added the pseudo_host_id to the original listing table for easier processing and removed it after completing the insertion of the data, in order to ensure compliance with the traceability dimension of data quality.

Due to foreign key dependencies, tables must be created in a valid order, in my case this meant starting with the Host table, then Listing and finally Rating. This is due to the fact that Host has no foreign key dependencies, Listing has a foreign key dependency on Host and Rating has a foreign key dependency on Listing.

First, I inserted the data into the Host entity, more specifically the pseudo_host_id, calculated_host_listings_count,, and host_since attributes from the listings table.

Next, I filled the Listing entity, inserting the listing_id (the pseudonomyzed IDs) from the Mapping.auto table, and pseudo_host_id, neighbourhood_cleansed, and price attributes from the listings table using a join statement.

Finally, I filled out the Rating entity, using the same listing_id as above from the Mapping.auto table and the number_of_reviews, number_of_reviews_ltm, review_scores_rating, review_scores_accuracy, review_scores_checkin, review_scores_cleanliness, review_scores_communication, review_scores_location, and review_scores_value attributes from

the listings table using a join statement. To filter out the entries with zero reviews I specified my requirements using the where statement.

Task 5

Single hosts versus multi hosts over the years

The first question we are tackling is the comparison of single hosts and multi hosts, based on their registration date and seeing if we can recognize the impact of the Coronavirus pandemic in the proportion between the two or the overall registrations in a year. All the queries for this question can be found in the 'Querying for question 1' tab in the DB Browser file.

Querying

To answer this question, few different queries were used, the first query seeks out the data on host registration grouped by year. To get this we used the 'STRFTIME' statement to extract the year from the host_since attribute, then the 'SUM' statement in combination with the 'CASE' statement was used to calculate the number of hosts with a single listing and the number of multi hosts. Finally using the 'COUNT' statement and the previous statements we calculated the percentage of multi hosts that registered every year, to receive the final result that line was also multiplied by 1.0. Finally, the 'GROUP BY' statement was used to separate the results by registration year.

The next query was just a simple look into the overall distribution of hosts, with a more detailed breakdown of multi hosts. The query for single hosts, multi hosts and the percentage of multi hosts remains the same as above without the 'GROUP BY' statement. To get the total number of a simple 'COUNT' statement counting all the rows in the host entity was used. Finally, the detailed distribution of the multi hosts used the same statements and attributes as the multi host and single host ones, just with the condition specifying the range of listings using the 'BETWEEN' statement.

The final query was used to compare the data before and after covid, 2020 was selected as the start of the post-covid era. The query again used the 'CASE' statement and the 'STRFTIME' statement to split the data into two groups, namely before 2020 and 2020 and after. The 'STRFTIME' statement separated the year from our host_since attribute and checked it against the set conditions to categorize it into a registration period. Then the same queries as above were used to calculate the total hosts, multi hosts and the percentage of multi hosts.

Analysis

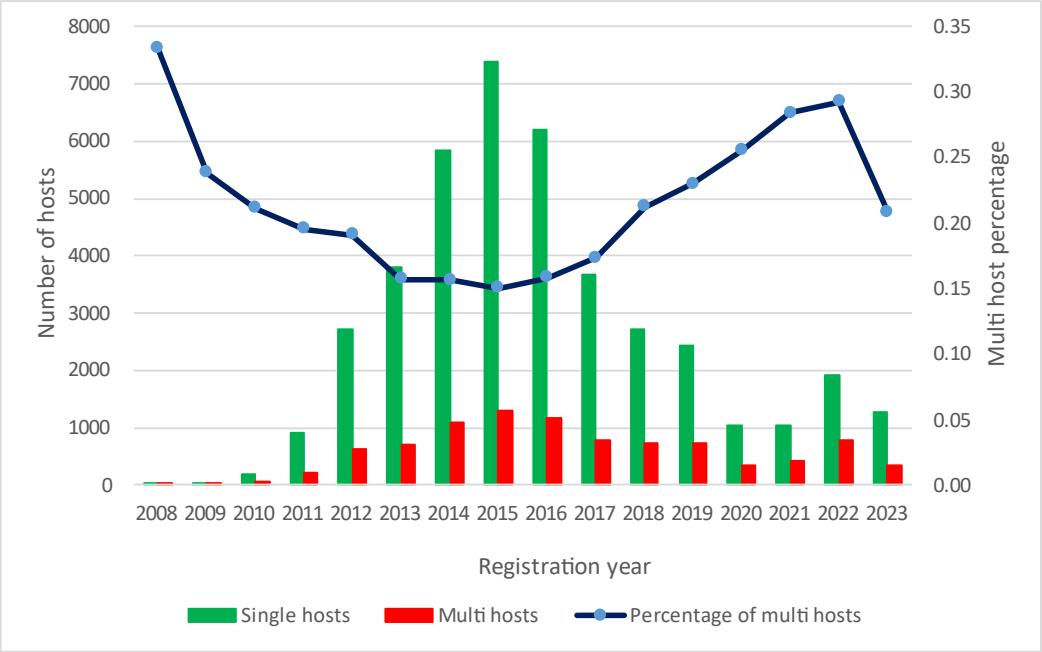

Figure 1 Number of hosts registered between 2008 and 2023

From the graph we can see that the number of registrations was steadily growing until 2015, where it peaked, after that registrations fell. We do not know the reason for the fall, but it could anything from stricter regulation impacting profitability and the attractiveness of being an Airbnb host, to just a simple saturation in the market. What we do know is that the number of single hosts registering fell more sharply than the number of multi hosts, which can also be seen from the steadily rising percentage of multi host registrations to total host registrations. After 2020 we can also see a slow pick up in number of registrations, but as 2023 is not over yet and the data not complete, we are not sure if this trend will continue.

Table 1 Distribution of hosts

Single hosts	Multi hosts	Total hosts	Percentage of multi hosts	2 - 10 listings	11 - 20 listings	21 to 30 listings	30+ listings
41178	9318	50496	0.18	8812	301	89	116

We can see that in total around 18% of all hosts on AirBnB are Multi hosts, of those the majority have between 2-10 listings, with only 5.4% of multi hosts having more than 10 listings.

Table 2 Multi hosts before and after 2020

Registration period	Total Hosts	Multi hosts	Percentage of multi hosts
Before 2020	43319	7418	0.17
2020 and after	7177	1900	0.26

Looking at the table we can see that after 2020 the percentage of multi-hosts among new registrants increased from 17% to 26%, which is a significant increase. Looking at the total number of hosts registered between the two periods is not very informative as the periods are not equal in length.

Overall, we can see that the number of new hosts has decreased since 2015, with a slight upwards trend since 2020. Out of those new hosts, however, more and more are multi-hosts, which might imply that Airbnb is becoming less accessible to new single hosts aiming to supplement their income, by renting out rooms or their house while away and is more orientated to towards multi-hosts who are using Airbnb to make a living.

Geographical distribution of listings

In the second question we are focusing on the geographical distribution of listings, and how it affects single host and multi host prevalence and a rough look at price.

Querying

All the queries can be found in the 'Querying for question 2' tab in the DB Browser file.

Here 3 queries in total were used to get all the information. All 3 queries used the 'WITH' statement to define a Common Table Expression. The first 2 queries looked at the neighbourhoods using the neighbourhood_cleansed attribute and the 'GROUP BY' statement, focusing on total listings per neighbourhood, number of listings operated by a single host, number of listings operated by a multi host, and the percentage of all listings in the neighbourhood operated by multi hosts. The first query ranked the neighbourhoods, using the 'RANK', 'OVER', and 'ORDER BY' statements, by percentage of all listings in the neighbourhood operated by multi hosts, while the second ranked them by total listings.

The second query was the same in structure only replacing the attributes with, average price, average price for listings operated by single hosts and average price for listings operated by multi hosts, ranked by average price.

In all cases the top 5 ranks were chosen, as those are the ones we are interested in, however the full tables are available in the appendix.

Analysis

Table 3 Top 5 Neighbourhoods by percentage of multi host listings

Neighbourhood	Total listings	Single host listings	Multi host listings	Percentage of multi hosts
City of London	463	129	334	0.72
Westminster	9329	2688	6641	0.71
Kensington and Chelsea	5381	1957	3424	0.64
Camden	5393	2236	3157	0.59
Hillingdon	801	333	468	0.58

Table 4 Top 5 Neighbourhoods by total listings

Neighbourhood	Total listings	Single host listings	Multi host listings	Percentage of multi hosts
Westminster	9329	2688	6641	0.71
Tower Hamlets	6633	3340	3293	0.5
Hackney	5828	3858	1970	0.34
Camden	5393	2236	3157	0.59
Kensington and Chelsea	5381	1957	3424	0.64

Table 5 Top 5 Neighbourhoods by average price

Neighbourhood	Average price	Average price single host	Average price multi host
Westminster	338.38	209.48	390.55
Kensington and Chelsea	327.19	262.88	363.95
City of London	246.68	195.71	266.37
Camden	212.31	170.38	242.01
Southwark	192.14	125.99	280.48

Our hypothesis that more 'touristy' and central neighbourhoods of London, would have a higher percentage of multi hosts appears to be correct, as the top 4 are in fact geographically located next to each other in the centre of London. City of London and Westminster are the most extreme examples with more than 70% of listings being operated by multi hosts.

In fact, looking at all the tables together, we can see that 3 of the highest multi host neighbourhoods are among the 5 biggest neighbourhoods by total listings and 4 of them are among the 5 most expensive neighbourhoods in terms of average price per listing. Additionally, Westminster tops all three tables, being the most expensive, the biggest in terms of listings and in terms of percentage of multi host listings.

Additionally, we can see from the price table that multi host listings are more expensive by quite the margin, with the single host listings all being below average in price.

Interestingly Hillingdon is the third least expensive neighbourhood in London, ranked by average price per listing.

Single hosts versus multi hosts on the rating scale

The final question is looks at the ratings grouped by whether the host is a multi-host or a single host. Is the personal touch appreciated by the customer or should the hospitality business be left to the professionals.

Querying

To answer this question, I used average ratings for every category split into two categories: single hosts and multi hosts. To compare the two, I used the 'AVG' and 'CASE' statements, to calculate the separate averages for single and multi-hosts. The 'CASE' statement was used to check if the host had 1 or more than 1 listing, then the 'AVG' statement was used to calculate the average of such cases.

Because the data came from all 3 of our entities, we had to join them all together, there was no problem in joining together the Host and Listing entities using the 'JOIN' statement, while for joining the Rating entity on top of that had to be done through a 'LEFT JOIN' statement as some listings did not have corresponding entries in the Rating entity. The query can be found in the 'Querying for question 3' tab in the DB Browser file. To remodel the table and produce the figure Excel was used.

Analysis

Figure 2 Bar plot showing the comparison in ratings between single and multi-hosts

Table 6 Table showing the comparison of single and multi-hosts in terms of reviews

Categories	Single host	Multi host
Average reviews	20.63	27.23
Average reviews in the last year	5.85	9.88
Average overall rating x/5	4.62	4.56
Accuracy x/5	4.67	4.62
Check-in x/5	4.71	4.71
Cleanliness x/5	4.55	4.56
Communication x/5	4.74	4.71
Location x/5	4.63	4.68
Value x/5	4.57	4.49

Looking at the number of reviews we can see that on average listings from multi-hosts receive more reviews than listings from single hosts, this is also seen in the number of reviews in the last year. Differences will be presented in parenthesis, with the calculation single host – multi host, unless stated otherwise.

From the table, we can see that single hosts have the advantage in most categories including the overall rating (0.06), barring Location (-0.05), where multi hosts edged out a slight advantage and Cleanliness (-0.01) and Check-in (0.00) where the difference is negligible. The categories in which single hosts have the advantage are Value and Accuracy, which matches up with the rationale that,

single hosts provide cheaper accommodation and are more thorough in describing the real situation, considering they are only focusing on one property.

The ratings are concentrated in the top end of the scale, meaning that people rarely give bad ratings, due to that the small deviations in ratings might be perceived as more meaningful, but even then, looking at the results we cannot conclusively say that the service is better in listings owned by single hosts compared to multi hosts.

Conclusion

Our findings suggest that while multi-hosts might be contributing to the centralization of market revenue and potentially exacerbating housing crises in certain areas, the disparity in service quality between single and multi-hosts, as evidenced by rating comparisons, is not starkly pronounced. The subtle advantages and disadvantages in various rating categories between the two host types indicate a nuanced customer experience, thereby not definitively leaning towards a single or multi-host superiority. Moreover, the geographical and pricing analyses hint towards a targeted operational approach by multi-hosts, potentially aligning their listings with touristic and high-demand neighbourhoods. Which could be further exacerbating the housing crisis in those high-demand neighbourhoods. Listings operated by multi-hosts seems to be more expensive on average compared to listings operated by single hosts, however there are many aspects of price we did not account for; therefore, our conclusions must be taken with a grain of salt.

In general, it seems that the trend of Airbnb is moving more and more towards multi-hosts, who are concentrating their listings in highly desirable areas charging higher than average prices, while providing a similar service. The results do lend credence to the theory that multi-hosts are contributing to the housing market crisis, but are not proof by themselves and such a claim would have to be corroborated with further research into the matter, taking a wider dataset not limited to Airbnb data.

Appendix

Table 7 Table of number of hosts registered per year

Registration year	Single hosts	Multi hosts	Total hosts	Percentage of multi hosts
2008	2	1	3	0.33
2009	32	10	42	0.24
2010	183	49	232	0.21
2011	906	220	1126	0.20
2012	2708	638	3346	0.19
2013	3798	708	4506	0.16
2014	5831	1081	6912	0.16
2015	7398	1309	8707	0.15
2016	6190	1164	7354	0.16
2017	3675	771	4446	0.17
2018	2731	736	3467	0.21
2019	2447	731	3178	0.23
2020	1037	355	1392	0.26
2021	1052	417	1469	0.28
2022	1914	792	2706	0.29

2023	1274	336	1610	0.21

Table 8 Ranking of neighbourhoods by percentage of multi host listings

Neighbourhood	Total listings	Single host listings	Multi host listings	Percentage of multi hosts	rank
City of London	463	129	334	0.72	1
Westminster	9329	2688	6641	0.71	2
Kensington and Chelsea	5381	1957	3424	0.64	3
Camden	5393	2236	3157	0.59	4
Hillingdon	801	333	468	0.58	5
Brent	2592	1144	1448	0.56	6
Barnet	2049	894	1155	0.56	6
Newham	2093	1009	1084	0.52	8
Havering	364	175	189	0.52	8
Ealing	2042	977	1065	0.52	8
Redbridge	787	386	401	0.51	11
Barking and Dagenham	524	257	267	0.51	11
Tower Hamlets	6633	3340	3293	0.5	13
Harrow	477	241	236	0.49	14
Croydon	1381	700	681	0.49	14
Bexley	453	230	223	0.49	14
Hammersmith and Fulham	3495	1812	1683	0.48	17
Hounslow	1116	588	528	0.47	18
Greenwich	1785	940	845	0.47	18
Enfield	748	398	350	0.47	18
Sutton	353	199	154	0.44	21
Southwark	4713	2695	2018	0.43	22
Islington	4442	2543	1899	0.43	22
Kingston upon Thames	635	374	261	0.41	24
Merton	1370	828	542	0.4	25
Wandsworth	4111	2516	1595	0.39	26
Lambeth	4432	2689	1743	0.39	26
Haringey	2201	1352	849	0.39	26
Bromley	703	438	265	0.38	29
Waltham Forest	1551	975	576	0.37	30
Lewisham	2383	1501	882	0.37	30
Hackney	5828	3858	1970	0.34	32
Richmond upon Thames	1158	776	382	0.33	33

Table 9 Neighbourhoods ranked by average price

Neighbourhood	Average price	Average price single host	Average price multi host	Rank
Westminster	338.38	209.48	390.55	1
Kensington and Chelsea	327.19	262.88	363.95	2
City of London	246.68	195.71	266.37	3
Camden	212.31	170.38	242.01	4
Southwark	192.14	125.99	280.48	5
Brent	186.25	135.9	226.03	6
Hammersmith and Fulham	184.65	150.84	221.05	7
Newham	184.16	124.2	239.98	8
Barnet	176.89	196.16	161.99	9
Wandsworth	176.54	148.67	220.5	10
Richmond upon Thames	170.17	161.17	188.45	11
Islington	166.06	127.7	217.42	12
Merton	158.05	147.76	173.78	13
Hounslow	155.68	109.71	206.88	14
Haringey	152.27	121.36	201.48	15
Lambeth	146.67	122.35	184.19	16
Redbridge	142.03	104.73	177.93	17
Kingston upon Thames	137.58	123.72	157.45	18
Tower Hamlets	132.95	116.11	150.04	19
Hackney	128.91	115.86	154.46	20
Ealing	126.44	119.48	132.83	21
Sutton	124.71	91.9	167.09	22
Greenwich	123.6	109.12	139.72	23
Havering	121.02	128.18	114.39	24
Waltham Forest	120.2	132.31	99.72	25
Lewisham	107.92	106.83	109.77	26
Bromley	105.27	100.13	113.77	27
Enfield	104.7	103.55	106	28
Barking and Dagenham	101.56	96.74	106.21	29
Harrow	100.11	90.37	110.06	30
Hillingdon	94.65	86.29	100.6	31
Croydon	94.48	91	98.04	32
Bexley	93.38	89.77	97.1	33

References

Norwood, G. (2022, January 13). Rents skyrocketing in London according to latest index. Landlord Today. https://www.landlordtoday.co.uk/breaking-news/2022/1/rents-skyrocketing-in-london-according-to-latest-index

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