Erasmus University Rotterdam

Rotterdam School of Management

Data Management & Ethics - BM02BAM Individual integrative assignment: The socio-economic impact of Airbnb in Paris

Gijs Werkman (615815)

Lecturers:	Anna Priante
Date:	9 October 2023

Contents

1. **Introduction** 1
    1.1. Context & Problem statement 1
    1.2. Research questions 1

2. **Design & organisation** 3
    2.1. Entities & attributes 3
    2.2. Relationships 4
    2.3. Normalisation 4

3. **Data processing** 5
    3.1. Checking data formats 5
    3.2. Renaming & creating variables 5
    3.3. Data deduplication 6
    3.4. Missing values 6
    3.5. Outlier detection 6
    3.6. Pseudonymisation 6

4. **Database implementation** 7

5. **Querying & reporting** 8
    5.1. Listings per neighbourhood 8
    5.2. Price of listings per neighbourhood 10
    5.3. Host registration patterns 12
    5.4. Distribution of listings per host 13
    5.5. Conclusion 14

A. **Appendices** 16
    A.1. Attribute overview 16
    A.2. SQL code data processing 16
    A.3. SQL code database implementation 19
    A.4. SQL code querying & reporting 20

1. Introduction

1.1. Context & Problem statement

In the dynamic landscape of urban tourism and the sharing economy, the City of Paris is facing complex challenges stemming from the rapid growth of peer-to-peer accommodation providers like Airbnb (Heo et al., 2019). This growth has sparked a multifaceted debate that resonates globally (Belarmino and Koh, 2020). While such platforms promise economic benefits and a novel sharing economy, they also induce socio-cultural, economic, and environmental repercussions that necessitate careful consideration (Hall et al., 2022). These impacts extend beyond the traditional hospitality sector, altering the flow of tourists and visitors, and reshaping the very fabric of neighbourhoods and communities (Hall et al., 2022).

Local governments seek to navigate this intricate terrain and make informed decisions regarding the impact of Airbnb on its communities, housing markets, and businesses (Adamiak, 2022). To effectively address these challenges, it is paramount for the City of Paris to gather insights into the geographic distribution of Airbnb listings, host behaviour, and the pricing dynamics across neighbourhoods. These insights will empower the city to develop strategic policies and interventions that balance the positive aspects of the sharing economy with the preservation of local culture (Oklevik et al., 2019), affordable housing (Ram and Tchetchik, 2022), and equitable economic opportunities (Törnberg, 2022).

1.2. Research questions

To gain a deeper understanding of these challenges, the following research questions will be explored:

1. Where are most Airbnb listings in Paris located?

This question identifies the geographic distribution of Airbnb listings throughout the city, shedding light on areas most affected by the presence of these listings. This can aid city officials in understanding how the presence of Airbnb is distributed across neighbourhoods and whether these are spatial uneven patterns similar to those observed in traditional hotels (Oklevik et al., 2019; Ram and Tchetchik, 2022), which holds significance for spatial planning and local development.

2. What is the distribution of daily price for Airbnb listings in each neighbourhood of Paris?

This question delves into the pricing dynamics across neighbourhoods, revealing variations in affordability and economic impact. Understanding these price distributions is imperative for addressing issues related to accessibility, affordability, and market competitiveness. Additionally, it is essential to explore how proximity to popular tourist hotspots may influence pricing, as neighbourhoods farther from these attractions may experience downward pressure on prices due to decreased demand (Tong and Gunter, 2022).

3. In which years did most hosts register on Airbnb?

This question explores historical registration trends among Airbnb hosts in Paris, offering insights into the platform's growth trajectory, including periods of rapid expansion and potential inflection points. This data is of interest to city planning, informing policymakers about evolving dynamics within the Airbnb ecosystem, market maturity, and potential impacts on the local housing market and communities.

4. What is the distribution of the number of listings manager per host in Paris?

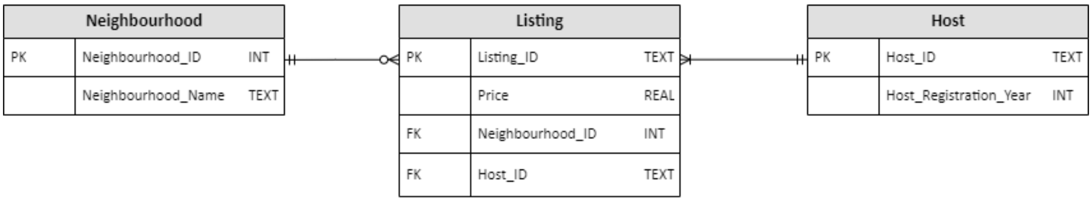
By examining how hosts engage with the platform in terms of the number of listings they manage, this question provides insights into the role of individual hosts in the Airbnb ecosystem. Moreover, it aids in gauging the level of platform utilisation among hosts and its implications for competition and market dynamics. Past research conducted by Törnberg (2022) has revealed that a considerable portion of the listings within the Airbnb marketplace is managed by a select group of hosts, emphasizing the concentration of activity among specific platform users. This underscores the potential influence on market dynamics and the pivotal role played by key actors in the Airbnb ecosystem.

By addressing these questions, the City of Paris is aided in formulating policies that balance tourism promotion, housing availability for residents, and the preservation of neighbourhood identity, contributing to the development of an equitable and sustainable urban environment in the era of the sharing economy.

2. Design & organisation

In light of the context, problem statements, and research questions delineated, the subsequent discussion will explore the data structures, database requisites, and the physical Entity-Relationship Diagram (ERD) depicted in figure 1. This ERD will serve as the foundational blueprint for the forthcoming database implementation.

Figure 1: Entity-Relationship Diagram (ERD)



Neighbourhood		
PK	Neighbourhood_ID	INT
	Neighbourhood_Name	TEXT

Listing		
PK	Listing_ID	TEXT
	Price	REAL
FK	Neighbourhood_ID	INT
FK	Host_ID	TEXT

Host		
PK	Host_ID	TEXT
	Host_Registration_Year	INT

2.1. Entities & attributes

To effectively address the research questions, careful consideration has been given to identifying the key entities and attributes. As part of the database design process, a naming convention has been adopted to enhance clarity. This convention has led to some changes in attribute names relative to the raw data, clarified in the attribute overview in appendix A.1.

To begin, the entity ‘Listing’ represents Airbnb listings in Paris, shaping the core of the analysis and relating to all research questions. ’Listing_ID’ serves as the unique identifier for each record and is the primary key. The ‘Price’ attribute signifies the daily cost of a listing, which is crucial for understanding affordability. ’Neighbourhood_ID’ and ’Host_ID’ function as foreign keys, establishing connections between listings, neighbourhoods, and hosts.

Moving on, the ’Neighbourhood’ entity plays a pivotal role in comprehending the geographic distribution of listings. This entity represents each of the arrondissements of Paris through a unique ‘Neighbourhood_ID’ tied to a ’Neighbourhood_Name’. This entity represents each of the arrondissements of Paris through ’Neighbourhood_ID,’ with ’Neighbourhood_Name’ providing essential contextual information. As the arrondissements of Paris are numbered from 1 to 20, the same numbering system will be used for ‘Neighbourhood_ID’.

Lastly, the entity ‘Host’ represents individual hosts on Airbnb. Each host is uniquely identified by a ‘Host_ID’, that being the primary key. Additionally, the ’Host_Registration_Year’ attribute has been added to denote the registration year of each host on the Airbnb platform. To prioritise privacy concerns and GDPR compliance, host names have been intentionally omitted

It is of worth noting that 'Host_ID' and 'Listing_ID' deviate from the conventional INT type typically assigned to unique identifiers in an ERD, as they are of type TEXT here. This modification is due to the implementation of pseudonymisation, where the original identifiers have been substituted with randomized hex values, consisting of both letters and numbers. This intentional alteration enhances privacy and security measures, fortifying the protection of sensitive data.

2.2. Relationships

The database design consists of the aforementioned entities, which are interconnected through their relationships, depicted in the ERD using Crow's Foot notation. This visual representation illustrates how the entities interact with one another, facilitating data analysis and supporting the exploration of the research questions.

Firstly, there is the relationship linking 'Listing' and 'Neighbourhood', which establishes a connection between individual Airbnb listings and the neighbourhoods in which they are situated. In this relationship, every Airbnb listing is associated with one neighbourhood. However, each neighbourhood may encompass multiple listings, but possibly also none at all.

Secondly, there is the relationship between 'Listing' and 'Host', illuminating the connection between Airbnb listings and their respective hosts. Here, a single host may manage multiple Airbnb listings, but each listing is ultimately associated with a single one host.

2.3. Normalisation

The provided ERD aligns with the principles of the first three normal forms (1NF, 2NF, and 3NF) for database normalisation. Each attribute in every entity contains atomic values, ensuring compliance with the 1NF requirement. To illustrate, consider the 'Host_Registration_Year' attribute, which has been thoughtfully introduced to succinctly represent the registration year of hosts. This modification streamlines previously composite data stored as 'date_since' in the DD-MM-YYYY format to YYYY, effectively adhering to the 1NF while furnishing essential information for analysis.

Additionally, adherence to 1NF means that duplicate rows are absent within the entities. Here, that is achieved through the use of primary keys. For instance, 'Host_ID' acts as a primary key in the 'Host' entity, ensuring the unique identification of each host record.

Moreover, the ERD enforces functional dependency of all non-key attributes on the entire primary key, solidifying its alignment with the 2NF requirement. For example, in the 'Host' entity, 'Host_Registration_Year' exhibits clear functional dependency on the primary key 'Host_ID.'

Furthermore, there is no presence of intermediary attributes that could cause transitive dependencies, ensuring compliance with the 3NF requirement. Thus, as the 2NF is met and transitive dependencies are avoided, the ERD satisfies the 3NF standard as well.

3. Data processing

To ensure sound decision-making and robust data analysis, it is imperative to conduct a thorough data quality assessment and cleansing process. This step is crucial for guaranteeing the accuracy, completeness, consistency, timeliness, traceability, and accessibility of the data at hand. Inadequate data quality can result in errors, misguided decision-making, and heightened operational inefficiencies. Therefore, identifying and rectifying any inconsistencies or inaccuracies within the data is of utmost importance.

For the Airbnb dataset, the data quality assessment and cleaning were performed using SQL. SQL was chosen for its exceptional data manipulation capabilities and its script-based nature, facilitating reproducibility and documentation. Since the research scope covers only a selection of the 75 attributes in the original dataset, the focus was placed on checking and cleaning those specific attributes. The SQL code employed for these data quality checks and cleaning procedures is readily accessible in appendix A.2. for reference.

To preserve the original data, a duplicate dataset, Listings_Original, was created. The cleaned version is stored in Listings_Cleaned.

3.1. Checking data formats

Upon importing the Airbnb dataset, the first step involved reviewing the data formats of all relevant attributes. This verification aimed to align the data with the specifications detailed in the ERD, depicted in Figure 1. Nearly all attributes were found to be consistent, with the exception of 'price.' To rectify this, currency symbols were removed, and the data type was converted from TEXT to REAL.

3.2. Renaming & creating variables

To allow for smoother interpretation and querying, the next phase entailed a comprehensive update of attribute names. This aligns with the predefined naming convention, as elucidated in Appendix 1. Furthermore, this stage involved the creation of two new attributes. Firstly, 'Host_Registration_Year' was generated based on information from 'host_since.' Secondly, 'Neighbourhood_ID' was introduced, derived from 'Neighbourhood_Name,' reflecting the arrondissement numbering system as previously stipulated.

3.3. Data deduplication

To ensure data integrity, an examination for duplicate records and listing IDs was executed. Fortunately, no duplicates were identified during this assessment. However, had duplicates been detected, a systematic process for removal would have been initiated. This not only safeguards the reliability of the dataset but also streamlines subsequent analyses by eliminating redundant information.

3.4. Missing values

Subsequently, a check was performed for missing values, identifying them solely in the attribute 'Host_Registration_Year'. Out of the 67942 records, 7 were found to have empty entries for this attribute. Rather than discarding these records entirely, a strategic decision was made to retain them. The rationale behind this choice lies in the recognition that the remaining information in these records is valuable. Nevertheless, it's important to note that these missing records will be excluded when delving into research question 3, which specifically examines the historical registration years of hosts.

3.5. Outlier detection

As part of the rigorous data cleaning process, attention was next turned to outlier detection. Here, the primary focus was on two key attributes: 'Price' and 'Host_Registration_Year'. It is worth noting that for the other attributes, whether they are unique identifiers or textual data, the concept of outliers is less applicable.

During the analysis of the 'Price' attribute, several potential outliers were identified through a systematic examination of the distribution. Outliers could significantly impact the understanding of the market and the factors influencing listing prices. However, instead of opting for outright removal, a more nuanced approach was chosen. Extreme values in pricing might indeed represent unique, high-value listings that enrich the dataset's diversity. By retaining these outliers, it is ensured that the full spectrum of the Airbnb market is reflected in the analysis.

Similarly, outlier detection was conducted for the 'Host_Registration_Year' attribute. Here, no extremely outliers were detected, with registration years ranging from 2008 to 2023.

3.6. Pseudonymisation

To safeguard sensitive information, the data cleaning process culminated with the implementation of pseudonymisation. A dedicated effort was made to replace identifiable details, specifically 'Host_ID' and 'Listing_ID,' with randomized hex values. This strategic measure, enacted through the creation of matching tables ('Matching_Table_Listing_ID' and 'Matching_Table_Host_ID') linking original and pseudonymised values. This fortifies the privacy and security of individual hosts and listings.

4. Database implementation

The database implementation followed the design outlined in the ERD in figure 1. This structured approach facilitates efficient data querying concerning specific listings, their associated neighbourhoods, and the respective hosts. All code used in the database implementation is available for reference in appendix A.3.

At the core of this database architecture is the ’Listings’ table, functioning as the primary entity. It incorporates a unique identifier for each listing, captures property prices, and establishes relationships through foreign keys pointing to the ’Neighbourhood’ and ’Host’ tables. These foreign keys not only ensure data integrity but also streamline querying processes by enabling seamless table joins for data retrieval.

The ’Neighbourhood’ table serves as a reference, containing unique identifiers and names for distinct neighbourhoods. This table provides neighbourhood information linked to each listing, eliminating redundancy through the use of distinct values and maintaining data accuracy.

Concurrently, the ’Host’ table tracks host information, featuring unique host identifiers and registration years. By isolating host details into a separate table, this design allows for possible future expansion of host-related information without the need for duplicating data across multiple listings.

During the population of these tables, the DISTINCT keyword was frequently used, ensuring that only unique values from the source data were inserted. This not only prevents data redundancy but also aligns coherently with the principles of normalisation, contributing to a well-organized and efficient database structure.

5. Querying & reporting

In this section, the results obtained through querying will be reported on. All SQL code used for these queries is available in Appendix A.4. Although the querying was solely done in R, some additional refining and visualisations were performed in R and tableau to provide the final output.

5.1. Listings per neighbourhood

1. Where are most Airbnb listings in Paris located?

To address this research question, a SQL query was employed, joining the ’Neighbourhood’ table with the ’Listings’ table using the ’Neighbourhood_ID’ as the common key. This allowed for the association of each Airbnb listing with its respective neighbourhood. The query utilised the COUNT (*) function to calculate the number of listings in each neighbourhood. Additionally, both the ’Neighbourhood_ID’ and ’Neighbourhood_Name’ attributes were included in the result set for reference. Finally, the results were sorted in descending order based on the number of listings, enabling the identification of neighbourhoods with the highest Airbnb listing counts.

As displayed in table 1, the neighbourhood of ”Buttes-Montmartre” boasts the highest number of listings, with a total of 7555 Airbnb listings. Conversely, the ”Louvre” neighbourhood exhibits the fewest listings, with only 1598 Airbnb listings.

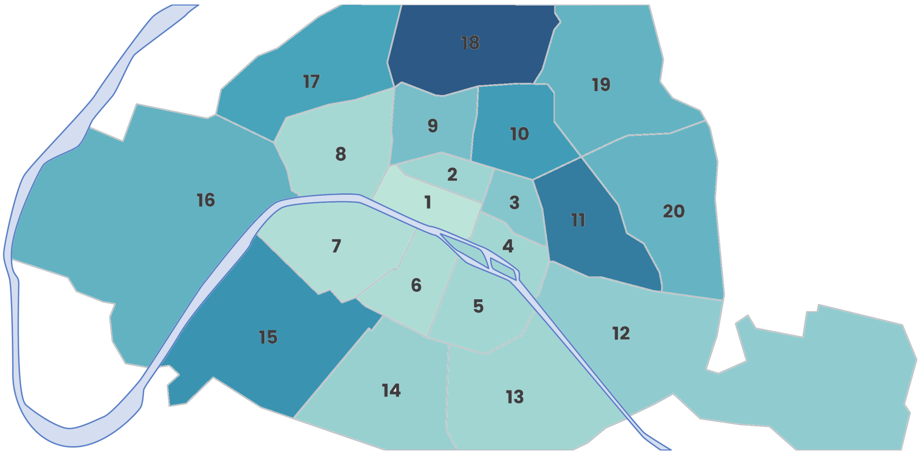
The spatial visualisation in figure 2 complements the tabular data, offering a clear geographical representation of listing concentrations. In this map, darker colours represent a higher count of listings, while lighter colours indicate fewer listings, and the numbers correspond to the Neighbourhood IDs

Strikingly, the majority of Airbnb listings are situated outside the city centre. This trend corresponds with the findings of previous research by Oklevik et al. (2019) and Ram and Tchetchik (2022), which suggest that hotels predominantly occupy city centres, while Airbnb listings tend to cluster on the outskirts. This phenomenon has implications for housing market dynamics and competition with local residents.

Table 1: Number of listings per neighbourhood

Neighbourhood ID	Neighbourhood	Number of Listings
18	Buttes-Montmartre	7555
11	Popincourt	6081
15	Vaugirard	5154
10	Entrepôt	4776
17	Batignolles-Monceau	4527
16	Passy	3914
19	Buttes-Chaumont	3866
20	Ménilmontant	3829
9	Opéra	3335
3	Temple	2993
12	Reuilly	2710
14	Observatoire	2489
2	Bourse	2337
13	Gobelins	2285
5	Panthéon	2243
4	Hôtel-de-Ville	2239
8	Élysée	2149
6	Luxembourg	1963
7	Palais-Bourbon	1899
1	Louvre	1598

Figure 2: Geographical spread of Airbnb listings



5.2. Price of listings per neighbourhood

2. What is the distribution of daily price for Airbnb listings in each neighbourhood of Paris?

To explore the distribution of daily prices for Airbnb listings across various neighbourhoods in Paris, a relatively simple, but efficient, SQL query was employed to extract the relevant data. The decision to use this query was driven by the need to obtain the daily price information for each listing within specific neighbourhoods, enabling a comprehensive analysis of price distributions across various regions of Paris.

The SQL query employed a JOIN statement to combine two tables: 'Listing' and 'Neighbourhood'. By joining these two tables using the 'Neighbourhood_ID' field, the query linked each Airbnb listing to its corresponding neighbourhood. The resulting dataset formed the basis for the subsequent analysis, allowing for the calculation of various summary statistics in R to understand the distribution of prices within each neighbourhood.

The analysis aimed to provide insights into how prices for Airbnb listings varied across different neighbourhoods in Paris. By utilising this dataset, summary statistics such as minimum price, quartiles (Q1 and Q3), median price (Q2), maximum price, and the count of outliers were computed for each neighbourhood. The presence of outliers was identified using the Tukey method, which relies on the interquartile range to detect values significantly different from the typical price range.

Notable findings include the substantial difference in median prices between neighbourhoods. In the upscale Élysée neighbourhood, the median price reaches €229, while in the more affordable Ménilmontant, it stands at just €90. This stark contrast underscores the vast range of affordability within the city, with implications for both residents and tourists seeking accommodation. Furthermore, the presence of extreme outliers, such as a maximum price of €63,594 in Palais-Bourbon, underscores the need for vigilant oversight and policy measures in the housing market to maintain a fair and accessible environment for both residents and visitors.

The observed variations in pricing dynamics underscore the importance of understanding affordability and accessibility, not only for tourists but also for the city's residents. The data in Table 2 demonstrates that different neighbourhoods may offer varying levels of affordability, potentially impacting the economic dynamics within the city. Additionally, it prompts a closer examination of how proximity to popular tourist attractions can influence pricing. As such, policymakers and stakeholders in Paris must consider these findings when crafting regulations and policies.

Table 2: Distribution of price per neighbourhood

Neighbourhood	Min	Q1	Median	Q3	Max	Outliers
Élysée	12	125	229	412	9999	179
Louvre	32	130	220	362.5	9970	106
Luxembourg	20	110	195	336.5	9999	151
Palais-Bourbon	10	110	194	336.5	63594	153
Bourse	10	113	180	300	9266	136
Hôtel-de-Ville	20	110	180	295.5	2800	146
Temple	9	107	179	280	11600	175
Passy	8	99.25	160	300	12000	349
Opéra	11	98.5	152	264	10770	183
Panthéon	10	94	150	253	5900	154
Vaugirard	13	80	120	199	11600	428
Entrepôt	20	80	120	190	10000	340
Batignolles-Monceau	15	78	118	197.5	9999	353
Popincourt	9	75	111	170	15000	441
Observatoire	15	71	105	165	8790	193
Buttes-Montmartre	8	70	103	150	21564	569
Reuilly	16	70	100	150	11600	222
Gobelins	10	65	99	150	10000	165
Buttes-Chaumont	10	64.25	94	131	30000	284
Ménilmontant	10	62	90	125	5000	267

5.3. Host registration patterns

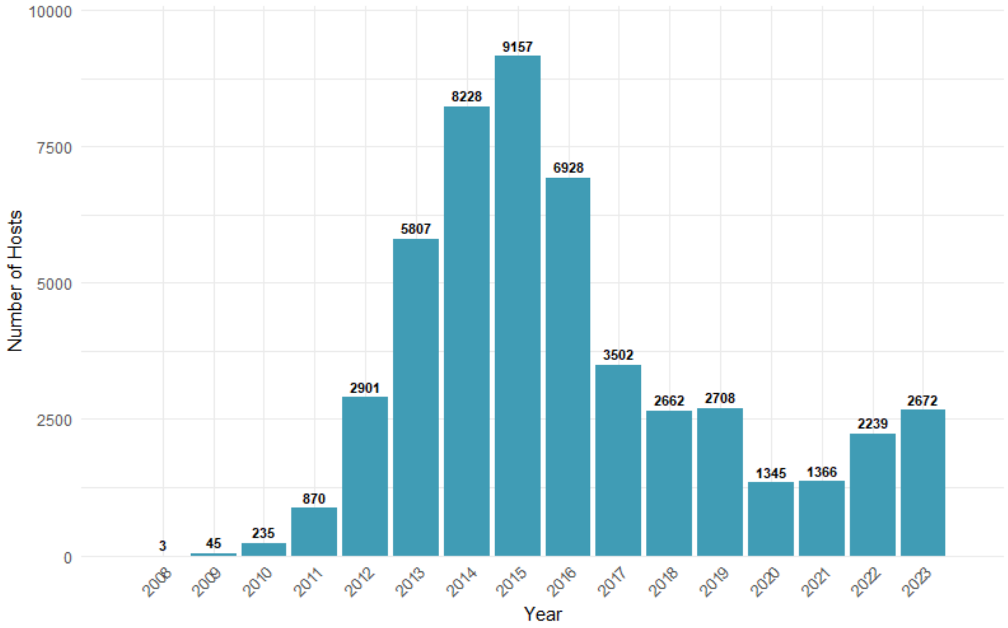
3. In which years did most hosts register on Airbnb?

To investigate the historical registration trends among Airbnb hosts in Paris, a SQL query that filtered out NULL values in the 'Host_Registration_Year' column of the 'Host' table was used. The data was grouped by registration year, and counts the number of hosts for each year. These null values were left out because, during the data processing phase, it was found that 7 empty entries existed for this attribute, which needed to be excluded for precise analysis.

The results, as depicted in Figure 3, offer a comprehensive historical overview of the registration patterns among Airbnb hosts in Paris. The data reveals that the earliest known host registration in an active capacity dates back to 2008. Over the years, the number of hosts exhibited a consistent upward trajectory, culminating in a peak in 2015 with 9157 hosts registering. However, subsequent years saw a decline in host registrations, with 2020 and 2021 recording lower numbers, potentially influenced by external factors such as the COVID-19 pandemic.

This recent decline in Airbnb host registrations carries significant implications for Paris. It suggests a potential transformation in the city's short-term rental landscape. Policymakers should seize this opportunity to comprehensively evaluate the impact of short-term rentals on the local housing market and communities. This assessment may prompt the need for a reevaluation of regulations to strike a delicate balance between the advantages of tourism and the preservation of affordable housing options for residents.

Figure 3: Registration years of active Airbnb hosts in Paris



5.4. Distribution of listings per host

4. What is the distribution of the number of listings manager per host in Paris?

To delve into the distribution of listings among Airbnb hosts in Paris, two SQL queries were employed to provide comprehensive insights into host engagement with the platform.

The initial query served the purpose of simply counting the number of listings managed by each individual host. It employed a straightforward SQL command that grouped the data by unique 'Host_ID' and tallied the occurrences of 'Listing_ID' associated with each host. An extensive dataset of 50674 rows was generated, with each row representing a different host alongside the count of listings under their management. This initial query was instrumental in laying the foundation for our analysis, but it presented a challenge due to the vast amount of data it produced.

Recognizing the need for a more organized and interpretable view of this data, a second query was developed to categorise hosts into different bins based on the number of listings they managed. This approach introduced structure to the distribution, making it more comprehensible. In this query, a CASE statement was utilised to allocate each host to a specific bin, ranging from '1' for those with a single listing to the '513-1024' bin for hosts managing a significant number of listings. Building upon the data generated by the first query, the second query facilitated a clearer understanding of how hosts engage with the platform.

The results of the second query are presented in table 3, which illustrates how hosts are distributed across various listing bins. Notably, the majority of hosts, totalling 46478 individuals, manage only one listing. As one moves to higher listing bins, the number of hosts progressively decreases, highlighting that hosts with multiple listings are less common. Interestingly, only three hosts in this dataset fall into the highest two bins, representing hosts managing over 256 listings. Cross-referencing this with the results from the first query verifies that the top three hosts indeed manage 262, 382 and 566 listings, respectively.

Understanding the distribution of listings among hosts carries significant implications for comprehending Airbnb's ecosystem dynamics. As alluded to in the introduction, previous research (Törnberg, 2022) has shown that a substantial proportion of Airbnb listings is concentrated among a select group of hosts. The insights revealed in table 3 suggest that while many hosts manage only one listing, there is also evidence of hosts overseeing numerous listings, possibly indicative of professional or commercial hosting activities. This disparity in host engagement underscores the necessity for a more nuanced examination of host behaviour and its impact on the Airbnb platform and the broader Parisian housing market.

Table 3: Distribution of Airbnb hosts by number of listings managed

Host count	1	2-4	5-8	9-16	17-32	33-64	65-128	129-256	257-512	513-1024
Listing bin	46478	3437	415	174	82	50	28	7	2	1

5.5. Conclusion

Based on the findings, several conclusions emerge.

Firstly, the spatial distribution of Airbnb listings is notably skewed towards the city's outskirts rather than its traditional tourist-centric areas. This spatial imbalance can impact local communities and businesses, making spatial planning and zoning regulations crucial for the city.

Secondly, pricing variations across neighbourhoods, from upscale areas like Élysée to more affordable ones like Ménilmontant, raise concerns about housing affordability for residents and market competitiveness. Proximity to tourist attractions further influences pricing, requiring policies that ensure affordability and fairness.

Thirdly, host registration trends highlight shifts in the short-term rental landscape, with a peak in 2015 followed by a decline. Policymakers should assess the impact of these changes on the housing market and communities.

Lastly, examining the distribution of listings among hosts uncovers disparities, with some managing multiple listings. This dynamic could affect market competition and community integration.

All in all, Paris faces challenges and opportunities in its relationship with Airbnb. Crafting strategic policies is essential to balance the benefits of short-term rentals with the preservation of local culture, affordable housing, and equitable economic opportunities. The insights provided may serve as a foundation for Paris to create a sustainable urban environment in the sharing economy era.

References

Adamiak, C. (2022). Current state and development of airbnb accommodation offer in 167 countries. *Current Issues in Tourism*, *25*(19), 3131–3149. https://doi.org/10.1080/13683500.2019.1696758

Belarmino, A., & Koh, Y. (2020). A critical review of research regarding peer-to-peer accommodations. *International Journal of Hospitality Management*, *84*, 102315. https://doi.org/https://doi.org/10.1016/j.ijhm.2019.05.011

Hall, C. M., Prayag, G., Safonov, A., Coles, T., Gössling, S., & Koupaei, S. N. (2022). Airbnb and the sharing economy. *Current Issues in Tourism*, *25*(19), 3057–3067. https://doi.org/10.1080/13683500.2022.2122418

Heo, C. Y., Blal, I., & Choi, M. (2019). What is happening in paris? airbnb, hotels, and the parisian market: A case study. *Tourism Management*, *70*, 78–88. https://doi.org/https://doi.org/10.1016/j.tourman.2018.04.003

Oklevik, O., Gössling, S., Hall, C. M., Jacobsen, J. K. S., Grøtte, I. P., & McCabe, S. (2019). Overtourism, optimisation, and destination performance indicators: A case study of activities in fjord norway. *Journal of Sustainable Tourism*, *27*(12), 1804–1824. https://doi.org/10.1080/09669582.2018.1533020

Ram, Y., & Tchetchik, A. (2022). Complementary or competitive? interrelationships between hotels, airbnb and housing in tel aviv, israel. *Current Issues in Tourism*, *25*(22), 3579–3590. https://doi.org/10.1080/13683500.2021.1978954

Tong, B., & Gunter, U. (2022). Hedonic pricing and the sharing economy: How profile characteristics affect airbnb accommodation prices in barcelona, madrid, and seville. *Current Issues in Tourism*, *25*(20), 3309–3328. https://doi.org/10.1080/13683500.2020.1718619

Törnberg, P. (2022). How sharing is the "sharing economy"? evidence from 97 airbnb markets. *PLOS ONE*, *17*, 1–19. https://doi.org/10.1371/journal.pone.0266998

A. Appendices

A.1. Attribute overview

Original attribute name	Updated attribute name	Type	Description	Notes
host_ID	Host_ID	TEXT	Airbnb's unique identifier for the host/user.	
	Host_Registration_Year	INT	Year in which the host/user created their account.	Newly created for this database using part of the from the original attribute 'host_since'.
id	Listing_ID	TEXT	Airbnb's unique identifier for the listing.	
	Neighbourhood_ID	INT	Unique identifier for the neighbourhood corresponding to the arrondissement.	Newly created for this database.
neighbourhood_cleansed	Neighbourhood_Name	TEXT	The neighbourhood as geocoded using latitude and longitude against neighbourhoods as defined by open or public digital shapefiles.	
price	Price	REAL	Daily price in euros.	

A.2. SQL code data processing

```
--Create a copy of the Listings table as Listings_Cleaned
CREATE TABLE Listings_Cleaned AS
SELECT *
FROM Listings;

--Rename the original Listings table to Listings_Original
ALTER TABLE Listings RENAME TO Listings_Original;

--3.1 Checking data formats
--Removing currency from Price
UPDATE Listings_Cleaned SET Price = CAST(REPLACE(Price, '$', '') AS REAL);
```

```
--3.2 Renaming & creating variables
--Updating attribute names
ALTER TABLE Listings_Cleaned RENAME COLUMN host_ID TO Host_ID;
ALTER TABLE Listings_Cleaned RENAME COLUMN id TO Listing_ID;
ALTER TABLE Listings_Cleaned RENAME COLUMN neighbourhood_cleansed TO Neighbourhood_Name;
ALTER TABLE Listings_Cleaned RENAME COLUMN price TO Price;

--Adding Neighbourhood_ID
ALTER TABLE Listings_Cleaned ADD COLUMN Neighbourhood_ID INTEGER;

UPDATE Listings_Cleaned
SET Neighbourhood_ID = CASE
    WHEN Neighbourhood_Name = 'Louvre' THEN 1
    WHEN Neighbourhood_Name = 'Bourse' THEN 2
WHEN Neighbourhood_Name = 'Temple' THEN 3
WHEN Neighbourhood_Name = 'Hôtel-de-Ville' THEN 4
WHEN Neighbourhood_Name = 'Panthéon' THEN 5
WHEN Neighbourhood_Name = 'Luxembourg' THEN 6
WHEN Neighbourhood_Name = 'Palais-Bourbon' THEN 7
WHEN Neighbourhood_Name = 'Élysée' THEN 8
WHEN Neighbourhood_Name = 'Opéra' THEN 9
WHEN Neighbourhood_Name = 'Entrepôt' THEN 10
WHEN Neighbourhood_Name = 'Popincourt' THEN 11
WHEN Neighbourhood_Name = 'Reuilly' THEN 12
WHEN Neighbourhood_Name = 'Gobelins' THEN 13
WHEN Neighbourhood_Name = 'Observatoire' THEN 14
WHEN Neighbourhood_Name = 'Vaugirard' THEN 15
WHEN Neighbourhood_Name = 'Passy' THEN 16
WHEN Neighbourhood_Name = 'Batignolles-Monceau' THEN 17
WHEN Neighbourhood_Name = 'Buttes-Montmartre' THEN 18
WHEN Neighbourhood_Name = 'Buttes-Chaumont' THEN 19
WHEN Neighbourhood_Name = 'Ménilmontant' THEN 20
    ELSE NULL
END;

--Adding Host_Registration_Year
ALTER TABLE Listings_Cleaned ADD COLUMN Host_Registration_Year INTEGER;
UPDATE Listings_Cleaned
SET Host_Registration_Year =
  CASE
    WHEN host_since IS NOT NULL AND host_since LIKE '____-__-__'
    THEN CAST(SUBSTR(host_since, 1, 4) AS INTEGER)
    ELSE NULL
  END;
```

```
--3.3 Data deduplication
--Check for duplicate records
SELECT listing_id, COUNT(*) AS DuplicateCount
FROM Listings_Cleaned
GROUP BY listing_id
HAVING COUNT(*) > 1;

--3.4 Missing values
SELECT * FROM Listings_Cleaned WHERE Host_ID IS NULL;
SELECT * FROM Listings_Cleaned WHERE Host_Registration_Year IS NULL;
SELECT * FROM Listings_Cleaned WHERE Listing_ID IS NULL;
SELECT * FROM Listings_Cleaned WHERE Neighbourhood_Name IS NULL;
SELECT * FROM Listings_Cleaned WHERE Neighbourhood_ID IS NULL;
SELECT * FROM Listings_Cleaned WHERE Price IS NULL;

--3.5 Outlier detection
--3.5.1 Outlier detection for Price
SELECT
  MAX(Price) AS max_price,
  MIN(Price) AS min_price,
  AVG(Price) AS avg_price
FROM Listings_Cleaned;

--3.5.2 Outlier detection for Host_Registration_Year
SELECT
  MAX(Host_Registration_Year) AS max_year,
  MIN(Host_Registration_Year) AS min_year
FROM Listings_Cleaned;

--3.6 Pseudonymisation
-- Matching table Listing_ID
CREATE TABLE Matching_Table_Listing_ID (
    Original_Listing_ID INT,
    Listing_ID TEXT
);

INSERT INTO Matching_Table_Listing_ID (Original_Listing_ID, Listing_ID)
SELECT DISTINCT
    Listing_ID AS Original_Listing_ID,
    hex(randomblob(16)) AS Listing_ID
FROM Listings_Cleaned;

-- Matching table Host_ID
CREATE TABLE Matching_Table_Host_ID (
    Original_Host_ID INT,
    Host_ID TEXT
);
```

```
INSERT INTO Matching_Table_Host_ID (Original_Host_ID, Host_ID)
SELECT DISTINCT
    Host_ID AS Original_Host_ID,
    hex(randomblob(16)) AS Host_ID
FROM Listings_Cleaned;

-- Update Listing_ID in Listings_Cleaned
UPDATE Listings_Cleaned
SET Listing_ID = (SELECT Listing_ID FROM Matching_Table_Listing_ID
WHERE Original_Listing_ID = Listings_Cleaned.Listing_ID);

-- Update Host_ID in Listings_Cleaned
UPDATE Listings_Cleaned
SET Host_ID = (SELECT Host_ID FROM Matching_Table_Host_ID
WHERE Original_Host_ID = Listings_Cleaned.Host_ID);
```

A.3. SQL code database implementation

```
--Create Listings entity
CREATE TABLE Listings (
    Listing_ID TEXT PRIMARY KEY,
    Price REAL,
    Neighbourhood_ID INTEGER,
    Host_ID TEXT,
    FOREIGN KEY (Neighbourhood_ID) REFERENCES Neighbourhood(Neighbourhood_ID),
    FOREIGN KEY (Host_ID) REFERENCES Host(Host_ID)
);

--Create Neighbourhood entity
CREATE TABLE Neighbourhood (
    Neighbourhood_ID INTEGER PRIMARY KEY,
    Neighbourhood_Name TEXT
);

--Create Host entity
CREATE TABLE Host (
    Host_ID TEXT PRIMARY KEY,
    Host_Registration_Year INTEGER
);

-- Populate Listings table
INSERT INTO Listings (Listing_ID, Price, Neighbourhood_ID, Host_ID)
SELECT DISTINCT
    Listing_ID,
    Price,
    Neighbourhood_ID,
    Host_ID
FROM Listings_Cleaned;
```

```
-- Populate Neighbourhood table
INSERT INTO Neighbourhood (Neighbourhood_ID, Neighbourhood_Name)
SELECT DISTINCT Neighbourhood_ID, Neighbourhood_Name
FROM Listings_Cleaned;

-- Populate Host table
INSERT INTO Host (Host_ID, Host_Registration_Year)
SELECT DISTINCT Host_ID, Host_Registration_Year
FROM Listings_Cleaned;
```

A.4. SQL code querying & reporting

```
--RQ1
--Count the number of listings in each neighborhood
SELECT Neighbourhood.Neighbourhood_ID, Neighbourhood_Name, COUNT(*) AS Number_of_Listings
FROM Neighbourhood
JOIN Listing ON Neighbourhood.Neighbourhood_ID = Listing.Neighbourhood_ID
GROUP BY Neighbourhood.Neighbourhood_ID, Neighbourhood_Name
ORDER BY Number_of_Listings DESC;

--RQ2
-- Select neighborhood and daily price data
SELECT Neighbourhood_Name, Price
FROM Listing
JOIN Neighbourhood
ON Listing.Neighbourhood_ID = Neighbourhood.Neighbourhood_ID;

--RQ3
--Count the number of hosts that registered in each year
SELECT Host_Registration_Year, COUNT(*) AS Number_of_Hosts
FROM Host
WHERE Host_Registration_Year IS NOT NULL
GROUP BY Host_Registration_Year
ORDER BY Host_Registration_Year;

--RQ4
--Count the number of listings managed by each host
SELECT Host_ID, COUNT(Listing_ID) AS Number_of_Listings
FROM Listing
GROUP BY Host_ID
ORDER BY Number_of_Listings DESC;
```

```
--Categorise the number of listings managed by each host into bins
SELECT
  CASE
    WHEN Number_of_Listings IS NULL THEN 'NULL'
    WHEN Number_of_Listings = 1 THEN '1'
    WHEN Number_of_Listings BETWEEN 2 AND 4 THEN '2-4'
    WHEN Number_of_Listings BETWEEN 5 AND 8 THEN '5-8'
    WHEN Number_of_Listings BETWEEN 9 AND 16 THEN '9-16'
    WHEN Number_of_Listings BETWEEN 17 AND 32 THEN '17-32'
    WHEN Number_of_Listings BETWEEN 33 AND 64 THEN '33-64'
    WHEN Number_of_Listings BETWEEN 65 AND 128 THEN '65-128'
    WHEN Number_of_Listings BETWEEN 129 AND 256 THEN '129-256'
    WHEN Number_of_Listings BETWEEN 257 AND 512 THEN '257-512'
    WHEN Number_of_Listings BETWEEN 513 AND 1024 THEN '513-1024'
    ELSE 'Over 512'
  END AS Listing_Bin,
  COUNT(*) AS Host_Count
FROM (
  SELECT Host_ID, COUNT(Listing_ID) AS Number_of_Listings
  FROM Listing
  GROUP BY Host_ID
) AS Subquery
GROUP BY Listing_Bin
ORDER BY MIN(Number_of_Listings);
```