BM02BAM - Data Management & Ethics

Individual Integrative Assignment

Data management for good: the social and economic impact of Airbnb

Athanasios Karavangelis - 674983

1. Introduction

The sharing economy, an economic model based on sharing underutilized assets from spaces to skills to stuff for monetary and non-monetary benefits (Botsman & Rogers, 2011) has transformed conventional paradigms of economic exchange in contemporary commerce and community. Valid examples of this model are the many short-term rental platforms, which have their roots in Couch-surfing - a network based on the ideals of sharing and non-monetary forms of exchange. Under the same panel, and although it's using monetary incentives, Airbnb, describes itself as part of this family of applications (Törnberg, 2022) and has evolved as a pioneer business in the world of short-term rentals by matching vacationers with unique housing options and playing an important role in the digital age transformation. Airbnb, claiming to be promoting peer-to-peer transactions and revealing the value of underutilized assets, seems to represent the promise of a more connected world. It makes the claim that it will assist with additional income (Levin, 2016) (Ganapavaram, 2022), strengthen local economies (Airbnb, 2021), and promote cross-cultural interaction (Airbnb, 2021), becoming known as a sign of innovation.

However, beneath the surface, an array of economic and social problems have emerged as a result of Airbnb's quick rise that demand further investigation. The platform's presence is complicated by issues like displacing tenants, encouraging gentrification, fostering over-tourism and overall a tourism industry vulnerable to negative externalities. These phenomena have led to several studies questioning whether AirBnB represents the sharing economy at all (Hall et al., 2022), as for example a study that shows that AirBnB has not only peer-to-peer but also business-to-clients offerings (Rodríguez et al., 2019), thus deviating from the core principles of the sharing economy. The problem is exacerbated as existing tenants are evicted and neighborhoods experience violent demographic changes as a result of Airbnb's tendency to close rent gaps in an unevenly geographic fashion (Wachsmuth, 2018). Moreover, Airbnb subtly allows "hosts" with multiple listings to leverage the platform to host their business (Levin, 2016), taking up many of the city's lodgings and leading to a push up on rents and an exclusion of the middle class (Cecco & Willsher, 2019), while of course causing negative reactions from local administrative forces (BBC, 2020).

Focusing more specifically on the side of the problem that can be illustrated through the use of data, in this analysis we investigate the city of Paris. According to Airdna, Paris is Airbnb's second-biggest market, only behind New York with the company hosting its annual home-sharing conference there in 2015 (Shih, 2015), its first occurrence outside of San Francisco, showing the importance of the French metropolitan city for the company. Nevertheless, the platform has caused an abundance of potential issues for the city, amongst which are gentrification, housing shortage and over-tourism to name some.

Our problem statement centers around these issues and is as following:
“Paris faces a housing shortage exacerbated by the proliferation of Airbnb listings catering primarily to tourists, totally altering the local landscape of central Paris and not only. WIth more and more houses offered for short-term rental through the platform, fueled by over-tourism, the rental gap in the city increases, especially in the central areas, causing radical increase in rent prices and gentrification. High-volume owners take advantage of these circumstances in an effort to maximize their profit through Airbnb, while parisian residents and minority owners seem to be slowly deteriorating in the platform.”

In order to examine this statement deeper, we will attempt to answer the following enumerated questions through results generated with the data at hand and proper illustrations and references that back up these results.

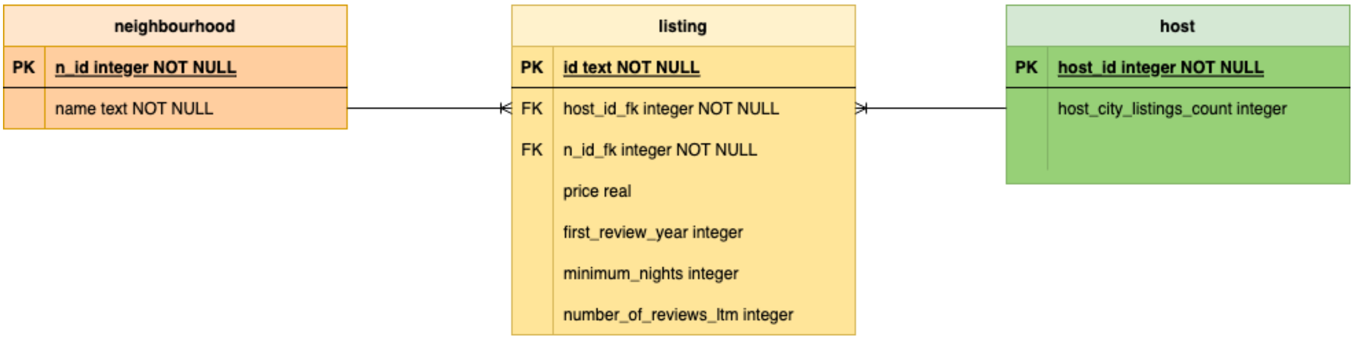
1. ***How has the listings number and average price changed in Paris throughout the years?***
   By answering this question we can roughly evaluate the effect of the presence of Airbnb in the last years in Paris, while also examining the listings from a first economic perspective, their price.
2. ***How can the hosts be categorized based on the number of listings they own and is the proportion amongst these host categories equal?***
   In this way, we will receive results regarding the number of hosts owning one or multiple listings in the city, and the respective number of listings that belong to each of these groups. This will shed light on potential inequalities existing regarding how the housing units are spread between the hosts.
3. ***How are listings spread across the many neighborhoods of Paris and is the listings distribution per neighbourhood similar per hosts’ category?***
   The distribution of listings across the so-called “arrondissements” of Paris can provide us an overview of the preferred areas for short-term rental offer and how this correlates with the tourist locations, landmarks and business areas of Paris. By having an additional overview of the listings location combined with the hosts category, we will be able to identify if there are different trending locations regarding each host category and examine the motives behind these trends.
4. ***What is the average price per night for Airbnb listings in Paris? How does this price vary across different arrondissements, and what distinctions exist between different hosts’ categories?***
   Analyzing the average price per night the listings, including variations across arrondissements and distinctions between single-listing hosts and multi-listing hosts for example, helps assess the economic impact of Airbnb on both hosts and guests. Also, it provides us insights into pricing strategies, affordability and existing distinctions between host categories.
5. ***What is the average number of booked nights per Airbnb listing in Paris? How does this metric differ across various arrondissements, and what disparities exist between different hosts’ categories?***
   Examining the average number of booked nights, along with differences between arrondissements and host types, reveals the preferences of the guests regarding location and time spent in the listings. It is important to state that the data insights from this question can highlight the popularity of Airbnb in different parts of the city and provide a first indication for which hosts benefit economically from their listings’ location.

6. ***What is the average yearly profit derived from Airbnb listings within the same arrondissement? How does this profit compare between single-listing hosts and multi-listing hosts?***
   Investigating the average yearly profit derived from Airbnb listings within the same arrondissement and comparing it between different host groups offers important insights into the financial benefits for hosts. Moreover, in this way we can assess the economic implications of hosting on Airbnb and the incentives of hosts to invest in the short-term rental market.

The research issues raised within the problem statement and questions are highly pertinent given the dynamic urban environment of Paris. The dynamics of Airbnb's presence in Paris must be understood by multiple stakeholders including governments, local communities and travelers alike. As it is a fact that multiple concerns about gentrification, housing shortage, malicious profit and the moral implications of short-term rentals have been sparked by the platform's effect on the local housing environment. In this assignment, we acknowledge the complexity of these challenges and aim to offer illustrative information that can, not necessarily lead to definitive answers but navigate towards fairer communities, adequate regulations ,and sustainable urban development.

2. Design & Organization

Regarding the steps we have to take in order to address the problem at hand, we have to first consider the modeling and organization of our data. For that reason, after iterating on the needs of our data problem, we create the following entity-relationship diagram which models best our data according to our questions and problem statement.



neighbourhood
PK n_id integer NOT NULL
name text NOT NULL
listing
PK id text NOT NULL
FK host_id_fk integer NOT NULL
FK n_id_fk integer NOT NULL
price real
first_review_year integer
minimum_nights integer
number_of_reviews_ltm integer
host
PK host_id integer NOT NULL
host_city_listings_count integer

Figure 1. Entity-relationship diagram of our database

In the design process of the diagram above, we followed several distinct steps. First, we had to consider what are the entities that are important for our database and apply best for our problem. The entities we opt for are *listing*, *host* and *neighbourhood.*
For the **neighbourhood** entity, we only need two attributes, an id, which we label *n_id* and is the primary key of our entity and the *name* of the neighbourhood. We need this entity in order to model the neighbourhoods (arrondissements) of Paris in an appropriate way, which are needed for the geographic interpretation of the problem and the examination of location trends regarding the listings.
As for the **host** entity, for our problem's needs we choose to include an id for the host, *host_id,* which serves as the primary key and the host's number of listings in the city of Paris, which we name as *host_city_listings_count.* This choice of ours can be attributed to the fact that in our problem we examine

the different groups of owners in the city and we want to focus on how they contribute towards existing issues. In this scope, it is of utmost importance to also include the number of listings each host owns in order to help us categorize the hosts in specific categories., which we will explain later on.
In the **listing** entity, we enclose all the needed information regarding an Airbnb listing that align with our problem which are its *id,* which serves as the primary key of the entity, its *price* (in dollars)*, first_review_year,* which indicates the year of the first review, *minimum_nights* per reservation and *number_of_reviews_ltm* which shows us how many reviews were made in the last year. It also includes two foreign keys *host_id_fk,* which corresponds to the *host_id* from the host table, and *n_id_fk,* which corresponds to the neighbourhood's *n_id.* The entity's structure was based on the fact that we look to delve deeper into the economic incentives that lie behind the ongoing housing situation in Paris, and thus we need to model the price of the listings combined with the number of reservations made for each listing. A problem with that was that our data do not include the exact number of reservations, and therefore we had to use an estimation of the reservations' number in our queries. The estimation method we use is based on the minimum nights per booking and the number of reviews a listing has in the past year. This method is based on Inside Airbnb's San Francisco model (Inside Airbnb, 2022), which we will explain further in the 5th section. Moreover, we include the *first_year_review* to help us model how the listings population in Paris and their price has changed over time.

Regarding **relationships and cardinalities** in our diagram, there is a one-to-many relationship between neighbourhood and listing, as one neighbourhood can be associated with multiple listings but the following cannot happen. Between host and listing there is once again a one to many relationship, given that a host can own multiple listings but the listing cannot have more than one host.

As far as **normalization** is concerned, in order to achieve 1NF, where our table has only atomic values, we split the *first_review* column of our data (where the date of the first review for the listing is stored) into *first_review_day*, *first_review_month* and *first_review_year.* All the other columns that we use contain atomic values and therefore in this way we have achieved 1NF. For the alleviation of transitive dependencies, and 2NF, we have to remove *host_id* and *neighbourhood_group_cleansed* from our data table because they are fully dependent on the listing's *id* column. To achieve that we create the *host* and *neighbourhood* tables , where we store *host_id* along with the *host_city_listings_count* and *neighbourhood_group_cleansed* as *name* along with *n_id* a column which we create and use as primary key. After that, we delete the problematic columns from our initial data table and therefore we achieve 2NF. To achieve 3NF, we need to ensure that there are no transitive dependencies. In our case, no transitive dependencies exist in our tables, and therefore we achieve 3NF.

Finally, there are some **GDPR** concerns which we have to consider regarding some of the attributes that we are using in our database. These lie in the listing's *id* and the *host_id*. As these attributes store values that are used in the actual Airbnb platform and can be used to link back to a specific real-world listing, we have to take measures in order to make sure that we do not violate the GDPR regulations. We explain how we tackled this issue in section 3.

3. Data Processing

In this section, we go through the data cleaning principles we applied, the tools we used to do so and the reasons behind each step of that procedure. Firstly, we should discuss the source and format of our data. Our data come from Inside Airbnb, a platform which openly shares specific data from many cities around the world regarding the homes listed on Airbnb. Our data is stored in a comma-separated values file and in this way we have to deal with only one large table, in terms of data processing, which includes all of our data.

The data processing steps we followed were based on the main principles for data quality, and we used an iterative process where we first check for each data quality principle and then take the necessary measures to achieve it.

First, we focus on the **accuracy** of data, where we check for mistakes in the data, duplicates and outliers. From a first check, our data does not contain any duplicate rows - i.e. listings - or duplicate values in the columns where there must be unique values, such as the id columns. Moreover, we check in the columns that contain dates if the specific date pattern is satisfied, using a pattern matcher in Python, and we validate that no violations were made in any rows. After that, we check for outliers in the *minimum_nights, number_of_reviews_ltm, number_of_reviews_l30d* and *price* columns. For minimum nights we use the interquartile method and specifically we remove values outside the range of the 5th percentile (0.05) to the 95th percentile (0.95), and in this way we eliminate extreme outliers that could distort the accuracy of our results. In addition, we remove rows where *number_of_reviews_ltm* exceeds 365 and *number_of_reviews_l30d* exceeds 30. This filtering eliminates cases where a property has an unusually high number of reviews within a short time frame or an unreasonably large total number of reviews. Lastly, we account for outliers in the *price,* where we remove the top 0.05% of the listings with the higher price. This method helps ensure that unusually high prices do not dominate and distort our results.

Secondly, we account for the **completeness** of the data, where we make sure that there is no missing data (e.g. null values). It must be stated that the data we use have been subject to a certain level of processing from Inside Airbnb and therefore we can argue that there is some level of completeness in the data, but from checks we do further on there are multiple columns with missing data. From these checks, we discover that there are missing values in multiple columns of the dataset that we are using and we therefore replace these missing values with appropriate values as detailed below:

- For **text** values, such as *description*, *host_about*, *neighbourhood_overview* we fill in appropriate text stating that no relevant information exists. In other cases such *neighbourhood* and *neighbourhood_cleansed* we aggregate information from *neighbourhood_group_cleansed* (as in the case of Paris all three correspond to the arrondissements), or we use the most commonly appearing value, we implement this in columns *host_acceptance_rate, host_response_time, host_response_rate.* Also, for many columns that contain text and have a binary meaning we make a convention and fill the missing values with the character ‘f’ in order to signal for ‘False’. We use this method for *host_is_superhost, host_has_profile_pic, host_identity_verified,* *license.*

- For **numeric** missing values we insert 0 where a null value would mean that there are probably no occurrences for this specific category. This is the case for the following columns: *minimum_minimum_nights*,*reviews_per_month, bathrooms,bedrooms, beds.* In other cases we calculate the average of the column, as for example in the 6 columns regarding review scores.

- For text values with **dates**, we use the value of the scraped date with the exception of the first_review column where we use the *host_since* column, as it is a more accurate way of estimating when the first review was made.

Next, in terms of the **consistency** of our data, where we check for format and other forms of inconsistencies, we do some processing in the columns of *price* and *license.* Column *price* was stored in our initial database as a string with the format being like '$100' and therefore this seems to be a wrong way of modeling for a price, which represents a number. Thus, we remove the '$' sign from the values and then restore the column with the *Real* data type which is used for real numbers. For column *license,* because there were many values, and because we use 'f' to state the non-existence of a license, we replaced each random value with one of the three values that Inside Airbnb uses, which are 't', 'Available with a mobility lease only ("bail mobilité")' and 'Exempt - hotel-type listing'.

In addition, in terms of the **traceability** of the data, and in order to alleviate GDPR concerns as stated in section 2 we had to take some measures regarding the listing's *id* and the *host_id.* What we did in order to anonymize these columns, and not be able to trace back to the original listings and hosts in the Airbnb platform was to create new columns with new unique id's, using 8 random digits for the new *host_id* and the first 3 letters of the neighbourhood followed by 10 random digits for the listing's *id*. For both of the columns we store a mapping of the new and old id's which we store in our database for validation purposes. The name of the mapping table for listings' *id* is ***listings_map*** and the name of the mapping table for *host_id* is ***hosts_map***. Regarding other columns which we would certainly not make use of such as *picture_url, host_thumbnail_url* and *host_picture_url,* we drop them in this stage to tackle any GDPR concerns.

Furthermore, regarding some other personal data that exist in our dataset that we will not use in our database, but for the sake of cleaning, we use **pseudonomization**. We use a Python package called *'Faker'* which can create realistic objects of names, company names, urls and others to pseudonomize the columns *host_name* and *host_url.* We also alter the *listing_url* in order for it to contain the new fake id we created for the listings.

Finally, a last step we execute for the correct structure of the *neighbourhood* table is that we create a mapping of the neighbourhood names (arrondissements) with the arrondissement numbers that correspond to the real numbers of the arrondissements in

Paris, in a table called *arrondissement_map*. In this way, we will use this mapping in the database creation to have the real arrondissement number as the *n_id*.

After all the steps of our data processing, we export the cleaned dataset in a new csv file, which we name *cleaned_listings.csv* and load it into our database into a table named ***cleaned_listings*** with caution for data type coherence. For all of our data processing we used Python and Jupyter Notebooks, along with *pandas* library which provides us a practical interface that helps us interact with our data quickly and efficiently.

4. Database Implementation

In this section, we will detail the process of implementing our database in accordance with the Entity-Relationship Diagram (ERD) developed in the 2nd section. We will describe the steps taken, explain the rationale behind our choices, and address any challenges encountered and their resolutions.

We initiated the database implementation by creating the *host* table, which includes two key attributes: *host_id* (as the primary key) and *host_city_listings_count*. Next, we designed the *neighbourhood* table which includes the *n_id* column,as the primary key, and the *name* column which stores the names of different arrondissements. We then create the central table of our database which is the listing table. This table includes the columns *id*, *host_id_fk* (a foreign key referencing the host table), *n_id_fk* (a foreign key referencing the neighbourhood table), *price*, *first_review_year*, *minimum_nights* and number_of_reviews_ltm.

To ensure that our database accurately reflects our dataset, we meticulously populated the tables using the appropriate SQL statements. The order of operations was crucial to avoid foreign key constraints and maintain data integrity. We began by populating the neighbourhood and host tables. These tables served as the foundation for the foreign key relationships that exist within the database. The SQL statements involved the selection of neighborhood *n_id*'s and *name* from the *arrondissement_map* we created in Section 3 and the extraction of the unique host information from the *cleaned_listings* dataset, using the DISTINCT SQL keyword. These steps ensured that our foreign keys could be appropriately mapped to their respective primary keys in the subsequent population of the listing table.

With the neighbourhood and host tables in place, we proceeded to populate the *listing* table using SQL statements to insert data from the cleaned dataset (*cleaned_listings*) and *neighbourhood* table into this table, using the JOIN statement based on the correct mapping of *neighbourhood_cleansed* values in cleaned_listing to the correct neighbourhood id's (*n_id_fk*) and host information to *host_id*'s (host_id_fk). This approach was crucial for maintaining data accuracy and ensuring the preservation of foreign key relationships.

The decisions made during the database implementation were driven by the need for data accuracy and adherence to the ERD. By creating and populating the *neighbourhood* and *host* tables first, we mitigated the risk of foreign key constraints when populating the listing table. In summary, the implementation of our database closely aligns with the ERD, ensuring that the database accurately reflects the dataset from Inside Airbnb.

5. Querying and reporting

In this section, we delve into the outcomes of our database exploration, meticulously structured to address each of the research questions posed earlier. Through a series of SQL queries, we seek to uncover valuable insights into the existing economic and social implications of the short-term rental market in Paris. The following subsections present our findings, each aligned with one of the questions, supplemented by reasoning for our query choices, diagrams, interpretations of results, and discussion of their implications.

1. How has the listings number and average price changed in Paris throughout the years?

To tackle this question, we began by creating a SQL query that provides a chronological view of the evolution of Airbnb listings in Paris. Our query leverages the power of the COUNT and AVG functions to calculate the number of listings and average price accordingly. The partitioning was done by year, using the GROUP BY statement and allowed us to split the data effectively and perform year-over-year comparisons. To ensure that our results are shown in a chronological progression we use the ORDER BY keyword based on *first_review_year.* Additionally, our query includes secondary nested subqueries that calculate the cumulative counts and average price of listings up to each year in our dataset. After running our query, we export the results into a CSV file to create an appropriate plot as shown below, which will explain the trends of number of listings and price in Paris.

The search results provided insight into how Paris's Airbnb scene has changed over time. There were only eight Airbnb listings in the city in 2009. But year after year, this number increased steadily, reaching a significant 61,663 listings by 2023. This increasing trend highlights Airbnb's fast-paced growth in Paris.

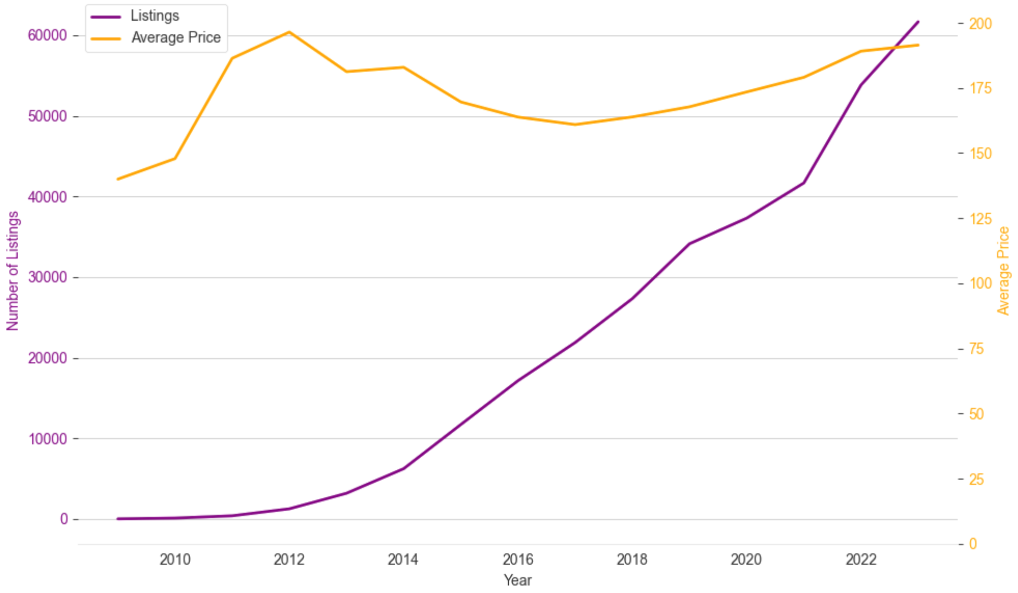


Figure 2. Yearly values of listings' number and average price ($) in Paris

Notable insights are also shown by the trend in the average price per night. The average daily rate in 2009 was 140$. We noticed price changes throughout the succeeding years. In 2012, there was a noticeable rise that peaked at 196$. But after that, prices showed a steady fall, and then a small increase to finally reach today's average price per night which is 192$.

We can say that our findings have significant implications for the Parisian short-term rental market. Airbnb accommodations are becoming more popular with tourists, giving hosts more opportunities to make money. Moreover, there are multiple concerns as the number of homes available for long-term residents may be limited as more homes are being turned into short-term rentals. The residents of the city may be impacted by this situation, which may also contribute to rising rents and a shortage of available housing. In this way, the rapid growth of Airbnb in Paris has the negative potential to reshape neighbourhoods and alter violently the demographic composition of certain areas, displacing long-term residents.

2. How can the hosts be categorized based on the number of listings they own and is the proportion amongst these host categories equal?

The SQL query employed for this analysis categorizes hosts based on the number of listings they own and provides relevant statistics. About the query's structure, we use a *CASE* statement, which acts as a conditional expression and categorizes hosts into our four groups: 'Single-listing Hosts,' '2-3 listing Hosts,' '4-9 listing Hosts,' and '10+ listing Hosts,' based on the number of listings each host owns. We then use the *COUNT* Function to count the number of hosts. We also calculate the total number of listings owned by hosts in each category using the *SUM* function for the *num_listings* column, which counts the number of listings owned by each host. Lastly, the query calculates two percentages: *percentage_of_listings* and *percentage_of_hosts*. These percentages are derived after using COUNT and *SUM* function and provide insights into the distribution of listings and hosts among the different categories. We also use the *ROUND* function to provide appropriately rounded numerical results, as well as the || operator to concatenate strings. The query uses *GROUP BY* to group the results by the hosts_group column, ensuring that statistics are presented for each host category and then orders the results in descending order of the *number_of_hosts* column using *ORDER BY DESC*.

To interpret our results in the best way possible we create a diagram that combines both the number of listings per host category (purple) and the number of hosts (orange) that constitute each of these categories. We also show the query results in a table format to allow for a more analytical overview.

hosts_group	number_of_hosts	total_listings	percentage_of_listings	percentage_of_hosts
Single-listing Hosts	42199	42199	68.43%	91.71%
2-3 listing Hosts	2893	6272	10.17%	6.29%
4-9 listing Hosts	646	3591	5.82%	1.4%
10+ listing Hosts	275	9601	15.57%	0.6%

Table 1. Results of the second query

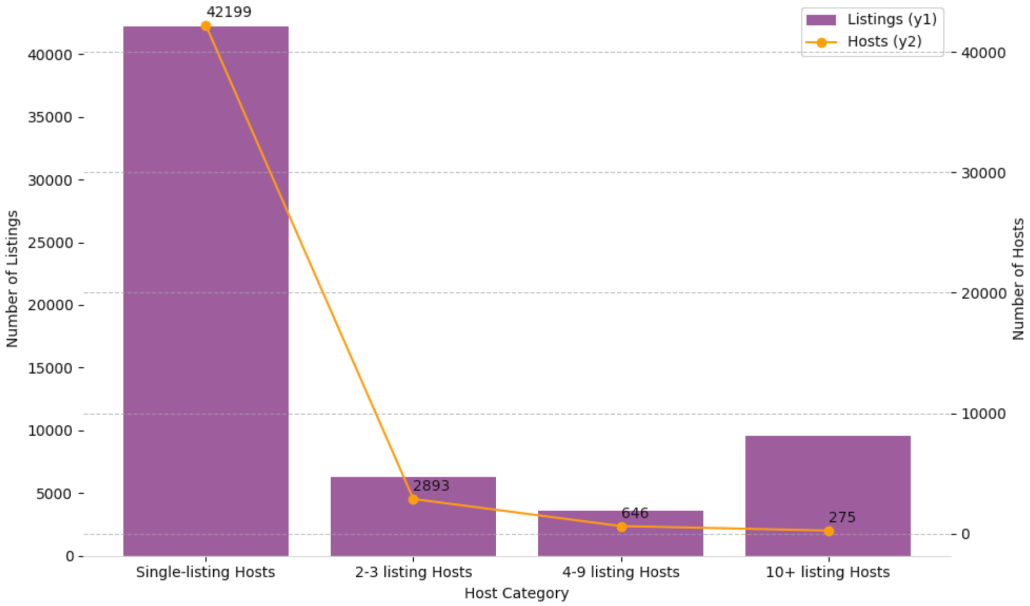


Figure 3. Bar plot of number of listings per host category with linear illustration of the number of hosts

Following the illustration of our results, one can point out that the majority of listings 68% are owned by single listing hosts which represent approximately 92% of the hosts. Surprisingly enough though, 15% of the listings are owned by 0.6% (!) of the hosts, which are the hosts with 10+ listings. This signals a huge economic inequality as a very big share of the listings belongs to few hosts and reflects a potentially different approach to short-term rental management, often associated with professional property managers, investors and entire businesses.

The distribution of hosts across various categories shows a first image of the hosts with 'Single-listing Hosts' dominating and '10+ listing Hosts' concentrating many listings. We assume that most single-listing hosts offer unique, diverse accommodations that align with sharing economy principles whereas the concentration of listings among few hosts suggests an economic focus, potentially impacting housing availability for long-term residents and necessitating specific regulatory approaches.

3. How are listings spread across the many neighborhoods of Paris and is the listings distribution per neighbourhood similar per hosts' category?

For this question we created two queries, one for each part of the question. In the first query, we use a straightforward *JOIN* operation that combines data from the *listing* and *neighbourhood* tables to count and identify the number of listings in each neighborhood. In the second query, we employ more advanced SQL techniques, including window functions such as *ROW_NUMBER()* and subqueries, to categorize hosts based on the number of listings they own and then assess the distribution of these host categories across neighborhoods. This query's complexity lies in the need to partition data into distinct host categories and rank neighborhoods accordingly too. Together, these queries offer valuable insights into the spatial distribution of listings and the role of different host categories in the short-term rental landscape in Paris.

In order to interpret the results of both queries we state the most popular arrondissements across all 20 of them and then we discuss how the host category influences the location of their listings. To do that, we create two maps of the Paris arrondissement, in the form of "heat" maps, where we color each arrondissement in a darker purple color in line with its ranking amongst all of them regarding the number of listings. To achieve that we utilized the *neighbourhoods.geojson* file of Paris from Inside Airbnb, which helps us map the arrondissement limits on the map, along with the specific results from our query and library *matplotlib* with Python.

To begin, according to the results of the first query, **the most popular arrondissements** amongst all categories of hosts are the **18th, 11th, 15th, 10th and 17th** in that order. These arrondissements as we can see in the diagrams below are located outside the center of Paris, with only one of the four central neighbourhoods of Paris being inside the top 10 arrondissements. However, the results of the second query, which are illustrated in the maps below, delve deeper into this analysis by categorizing hosts based on the number of listings they own and examining the distribution of these host categories across neighborhoods. This approach allows us to identify whether different host categories prefer specific locations and provides insights into the motives driving these trends.

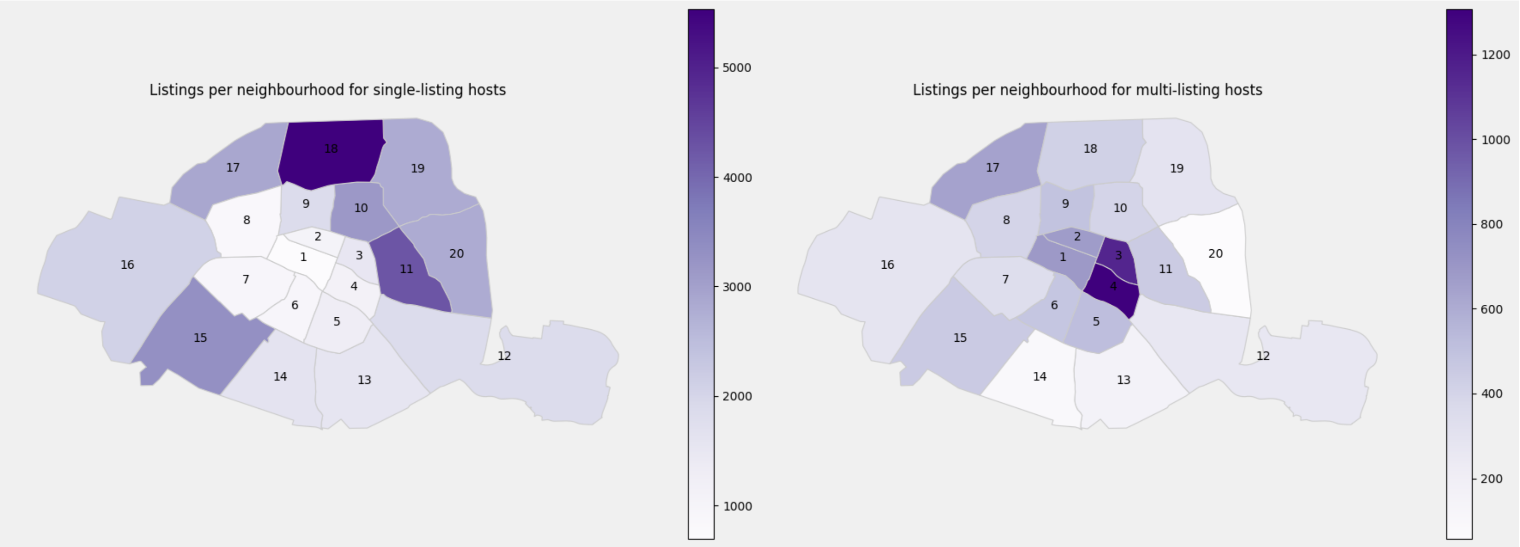


Figure 4. Maps of the city of Paris which illustrate neighbourhoods with a higher number of listings with darker accents of purple, for single-listing hosts(left) and multi-listing hosts(right).

For hosts with only one listing, we see that the city's center, where many popular tourist attractions, landmarks, and commercial hubs are located, appears whiter, indicating fewer listings. However, the outer ring surrounding the city center is shown as purple in contrast, which represents a higher density of listings. In contrast, the heatmap **for hosts with multiple listings**, paints a different picture. Here, the city's center is highlighted in purple to denote a greater number of listings. We can see that as we move away from the city center the neighbourhoods have a lighter color, indicating less listings. This suggests that multi-listing hosts own and are more inclined to invest in properties closer to the heart of Paris.

The excessive presence of Airbnb properties in the central area can result in the commercialization of the residential neighborhoods in the center of Paris. The influx of tourists and transient visitors can change the character of these areas, prioritizing the needs and interests of tourists over those of local residents. This can also lead to the displacement of long standing local businesses, with new businesses caring only for the preferences and disposable income of wealthier newcomers. Lastly, in neighborhoods with a high concentration of multi-listing hosts, a significant portion of housing stock is converted into short-term rentals, thus reducing the availability of housing for permanent residents and making the problem of housing shortage more serious.

4. What is the average price per night for Airbnb listings in Paris? How does this price vary across different arrondissements, and what distinctions exist between different hosts' categories?

For this question where we examine the average price of listings in Paris through various aspects we create **four queries** which address each separate part of our question. In the first query, we use the *AVG* function to calculate the overall average price per night for all listings in Paris. The second query utilizes JOIN operations to combine data from the *'listing'* and *'neighborhood'* tables, grouping the results by neighborhood. This query identifies and ranks the neighborhoods based on their average price per night, offering insights into pricing disparities across different areas of Paris. The third query introduces a more complex SQL syntax, including subqueries, to categorize hosts based on the number of listings they own and calculate the average price per night for each host category using *AVG* with a *PARTITION BY* statement. This query partitions data by host category and provides a deeper understanding of how listing prices vary based on host groups. Lastly, the fourth query extends the complexity by combining neighborhood data with host category data to determine the average price per night for each host category in each neighborhood. We use a subquery and the *CASE* statement for the hosts group and a *GROUP BY* statement for hosts group and neighbourhood to reveal the relationship between host categories, neighborhood locations, and pricing trends.

Our results show that the average price per night for Airbnb listings in Paris is approximately **192$**. However, this does not show much, as there are significant price variations among neighborhoods, with the **Élysée(8th arrondissement)** neighbourhood having the highest average price per night (382$) and Ménilmontant (20th arrondissement) having the lowest (109$). It is important to note that 7 of the 8 most expensive neighbourhoods in terms of average price, belong in the 8 most central areas of Paris signaling for a high correlation between high pricing and central location. This has a serious implication regarding hosts categories as multi-listing hosts own most of their listings in the central neighbourhoods and thus are imminent for much larger profits than other hosts categories which own most of their listings in less central neighbourhoods. The results and conclusions we discuss are adequately visualized in the map below which illustrates which are the most expensive arrondissements of Paris.

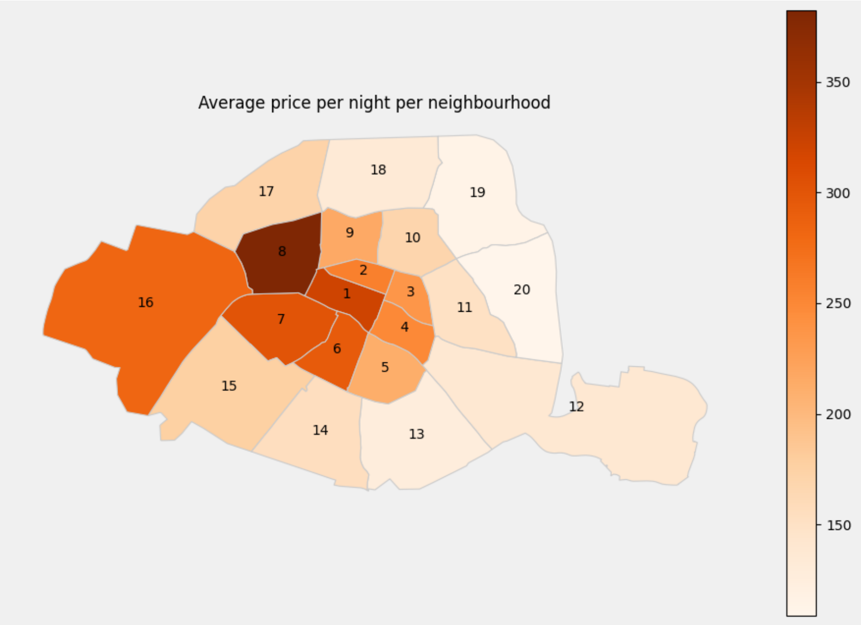
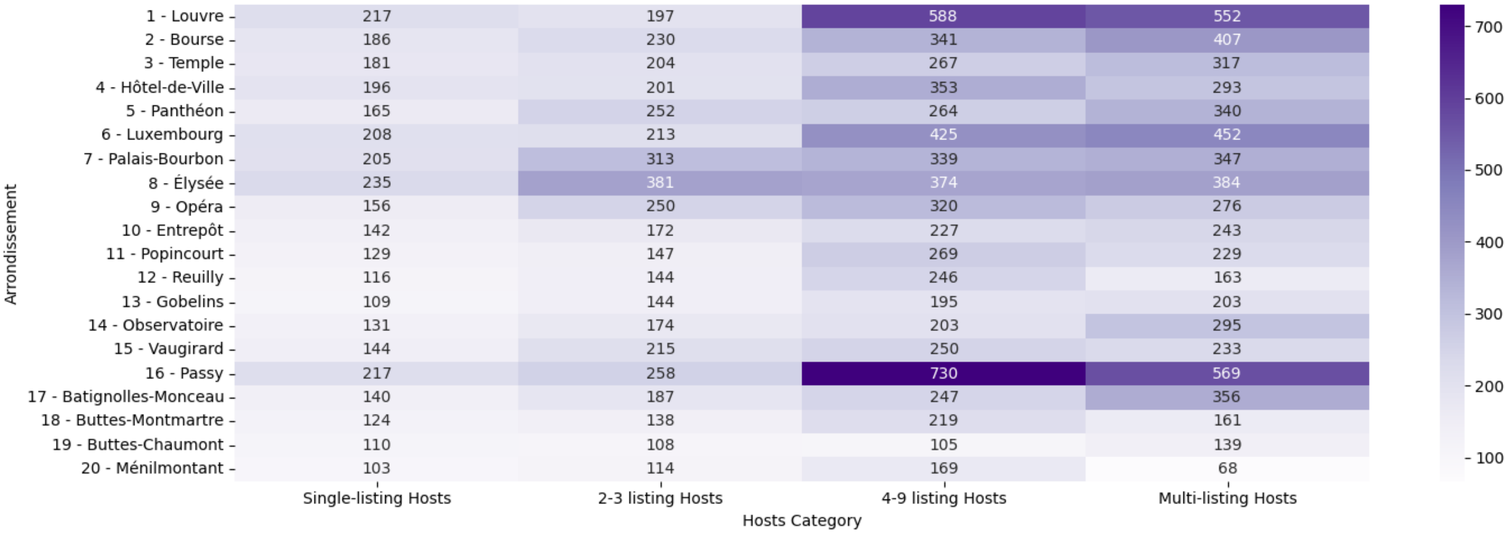


Figure 5. *Map of the city of Paris which illustrates the neighbourhoods based on average price per night*

Moreover, when we examine pricing amongst the different host categories we can validate multiple inequalities with multi-listing hosts having the highest average price (200$), followed by 2-3 listing hosts (130$), single-listing hosts (106$), and 4-9 listing hosts (70$). This along with the results of the fourth query, shown in Figure 6 below, where we have an analytical image from every neighbourhood, shows that multi-listing hosts dominate pricing in most neighborhoods, with their listings commanding largely higher prices compared to other host categories. This dominance varies across neighborhoods but generally follows a consistent pattern where multi-listing hosts have the highest average prices.



Arrondissement	Single-listing Hosts	2-3 listing Hosts	4-9 listing Hosts	Multi-listing Hosts
1 - Louvre	217	197	588	552
2 - Bourse	186	230	341	407
3 - Temple	181	204	267	317
4 - Hôtel-de-Ville	196	201	353	293
5 - Panthéon	165	252	264	340
6 - Luxembourg	208	213	425	452
7 - Palais-Bourbon	205	313	339	347
8 - Élysée	235	381	374	384
9 - Opéra	156	250	320	276
10 - Entrepôt	142	172	227	243
11 - Popincourt	129	147	269	229
12 - Reuilly	116	144	246	163
13 - Gobelins	109	144	195	203
14 - Observatoire	131	174	203	295
15 - Vaugirard	144	215	250	233
16 - Passy	217	258	730	569
17 - Batignolles-Monceau	140	187	247	356
18 - Buttes-Montmartre	124	138	219	161
19 - Buttes-Chaumont	110	108	105	139
20 - Ménilmontant	103	114	169	68

Figure 6. *Heat map that highlights the most expensive combinations of neighbourhood and host category*

The pricing disparities that we observe, particularly driven by multi-listing hosts, can have a significant impact on housing affordability and may necessitate regulatory changes. The conversion of residential properties into short-term rentals for the sake of larger profit can reduce the availability of long-term rental units in central neighborhoods. This scarcity can result in housing shortage and largely increased rent prices for local residents. To combat these effects, regulatory measures may be required, such as zoning restrictions, occupancy limits, or taxation policies, aimed at striking a balance between accommodating tourists and ensuring housing options for residents.

5. What is the average number of booked nights per Airbnb listing in Paris? How does this metric differ across various arrondissements, and what disparities exist between different hosts' categories?

To answer this question, we use **three separate queries** which can help us identify the booking patterns using overall averages, host categories, and neighborhoods. For this question and question 6, we need to calculate the number of booked nights. To do so, we use **Inside Airbnb's San Francisco model** (Inside Airbnb, 2022), which employs a **review rate of 50%** to convert reviews into estimated reservations, thus one review equals two reservations. To calculate the number of average booked nights per reservation we use the **average booked nights for Paris in 2023 which are 4.7** based on Inside Airbnb's calculations and results. In cases where a listing specifies a minimum nights requirement that exceeds the Paris average length of stay, we prioritize the minimum nights value for our calculations. It is important to say that we only include listings that have had one listing in the last 12 months and thus qualify as "active" listings.

The first query aims to calculate the overall average number of booked nights per listing in Paris. To achieve this, it uses the *AVG* function to compute the average of the "price" column. This column represents the price per night for each listing. The second query focuses on understanding how booking patterns differ among various host categories, such as single-listing hosts and multi-listing hosts. Within this query, we use a subquery which utilizes the *CASE* statement to categorize hosts into different groups. The main query then groups the data by host category using the *GROUP BY* clause and calculates the average booked nights for each category by applying the *AVG* formula. The third query is designed to rank neighborhoods based on the average number of booked nights for listings within each neighborhood. Similar to the previous query, a subquery is used to calculate the length of stay and select relevant columns from both the *listing* and *neighbourhood* tables. The main query aggregates the data by neighborhood using the *GROUP BY* clause and computes the average booked nights for each neighbourhood with the *AVG* function. To facilitate ranking, the *ROW_NUMBER* function is applied to assign a rank to each neighborhood based on their calculated average booked nights.

To our results, from the first query we find that the active listings in Paris (have a review in the last year) are 30762 - almost half - and are booked at an **estimated average of 107 nights per year**. Regarding the second and third queries, we illustrate their results using a table and a barplot accordingly which help us visualize the existing disparities amongst host groups and how this intensifies based on the neighbourhood.

hosts_group	average_booked_nights	number_of_listings
10+ listing Hosts	119.86	5854
2-3 listing Hosts	118.88	3166
4-9 listing Hosts	106.73	1869
Single-listing Hosts	101.48	19873

Table 2. The results of the second query, showing the average booked nights per hosts group

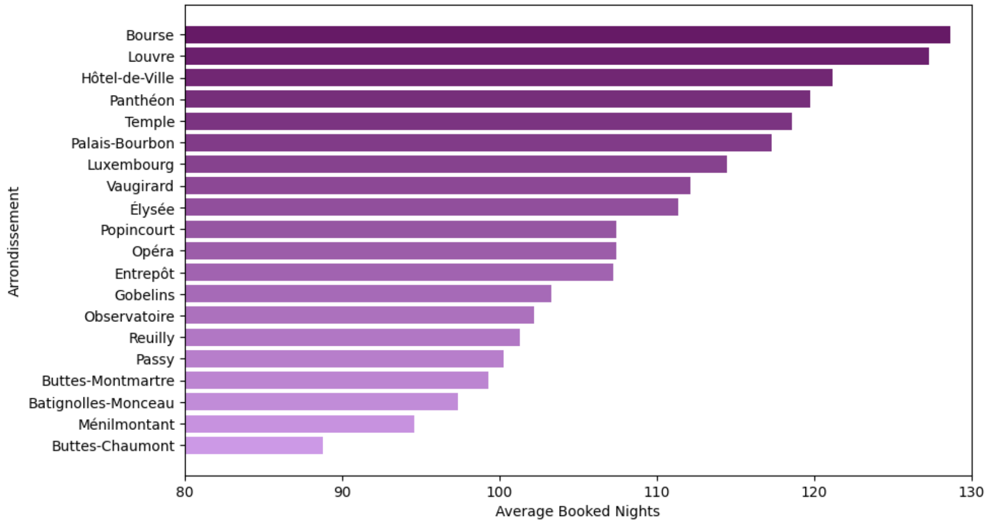


Figure 7. Bar plot that depicts the average booked nights per arrondissement

To interpret the results from this query in the best way possible, we will combine the implications arising with the results of question 6 which sheds light on hosts income per listing.

6. What is the average yearly profit derived from Airbnb listings within the same arrondissement? How does this profit compare between single-listing hosts and multi-listing hosts?

The three queries we formulate for question 6 are designed to calculate the average yearly profit derived from Airbnb listings and to compare this profit between single-listing hosts and multi-listing hosts. The first query calculates the average yearly income for hosts in each category, considering the number of reviews, length of stay, and listing price. It groups the results by host categories and provides the average yearly income along with the count of listings in each category in a similar structure as the 1st query in question 5. The second query calculates the estimated average yearly income for listings in each arrondissement and ranks the results by yearly income. The final query takes our analysis a step further and refines the analysis by calculating the estimated average yearly income for listings in each arrondissement for all host categories. The structure for all the queries is similar to the one used in question 4 and 5 and in order to not become repetitive, we do not detail them further.

The results of our queries are described and discussed below, but what must be given specific importance is the diagram below which shows that 10+ listing hosts make almost three-times(!) as much profit per listing as the single listing hosts. This leads to an enormous disparity as this difference in the income per listing and not collectively is translated to a disproportionate share of the short-term rental market for the multi-listing hosts compared to that of the single listing hosts.

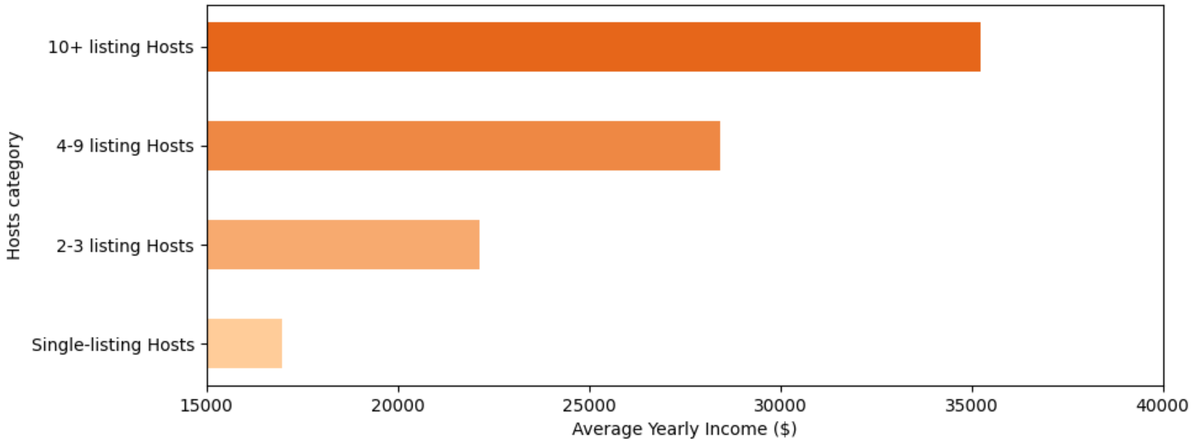


Figure 8. Bar plot that depicts the estimated average yearly income per listing per host category

The analysis of the average yearly profit from Airbnb listings within the same arrondissement, while comparing it across different host groups, provides crucial insights into the financial gains that hosts can expect. This examination also allows us to gauge the economic implications of participating in Airbnb hosting and the motivations driving hosts to engage in the short-term rental market. Our results reveal that hosts who manage multiple listings experience more substantial financial benefits as the combination of higher number of average booked nights and higher rates in more central locations of Paris leads to increased revenue. Given that these hosts probably take advantage of economies of scale in their property management, we are led to the conclusion that there is a huge gap between this king of business-oriented hosts and the sharing economy that Airbnb claims to be an example of.

Looking into specific arrondissements in Table, it becomes evident that certain central neighbourhoods like the Louvre, Bourse, and Élysée offer particularly attractive financial incentives for hosts. These central locations that benefit from higher demand, enable hosts to command higher prices and ultimately result in greater yearly income. On the other hand, arrondissements that are popular amongst single-listing hosts generate less average income per listing, causing greater inequality.

ranking	n_id	name	average_yearly_income	number_of_listings
1	1	Louvre	41850	765
2	2	Bourse	36801	1244
3	8	Élysée	35504	985
4	6	Luxembourg	32892	979
5	7	Palais-Bourbon	32273	833
6	3	Temple	30602	1562
7	4	Hôtel-de-Ville	30429	1165
8	5	Panthéon	26726	1077
9	16	Passy	23437	1653

10	9	Opéra	23428	1491
11	15	Vaugirard	21442	2181
12	10	Entrepôt	20277	2272
13	11	Popincourt	18196	2740
14	17	Batignolles-Monceau	16799	1820
15	14	Observatoire	15690	1112
16	12	Reuilly	15477	1134
17	18	Buttes-Montmartre	14671	3441
18	13	Gobelins	14198	974
19	20	Ménilmontant	11029	1660
20	19	Buttes-Chaumont	10984	1674

Table 3. The results of the the estimated average yearly income per neighbourghood

Finally, the breakdown of average yearly income by both arrondissement and host category illustrates once again the nuanced financial landscape of Airbnb hosting in Paris. For example, in all of the central arrondissements, multi-listing hosts tend to achieve the highest yearly income, emphasizing the role of location and host category in income disparities.

6. Conclusion

To conclude, this comprehensive analysis of Airbnb listings in Paris has generated valuable insights into the existing situation regarding the short-term rental market in this iconic city. Through a series of well-structured SQL queries and data-driven examinations, we've uncovered key patterns, such as uneven spatial distribution, host categories, pricing, occupancy rates, and profitability. One can argue that our findings underscore the presence of geographical disparities, with central arrondissements commanding higher prices and yielding immense profit. Multi-listing hosts, that take advantage of the platform for the profit possibilities, stand out as they seem to be the host group that creates imbalance and is the root of inequality. Furthermore, our investigation into pricing and occupancy rates has unveiled critical economic implications for hosts, neighbourhoods, and policymakers, highlighting variations in profitability across different host categories and locations. These insights suggest that the short-term rental market highly contributes to the existing issues of gentrification and housing shortage in Paris, as the hosts have a profit-driven approach to property management which leads to less housing supply for Parisian residents and disruption of local communities. In essence, this project demonstrates the highly important role of data-driven analyses in understanding the particularities and hidden patterns behind the false type of sharing economy that Airbnb exemplifies. This study paves the way for further research and policy discourse in the realm of short-term rentals, emphasizing the significance of data-driven insights in resolving and contemplating the existing issues and disparities in today's world.

References

Airbnb. (2021, August 9). *Airbnb announces partnership with UNESCO to promote cultural tourism*. Airbnb Newsroom. Retrieved October 9, 2023, from https://news.airbnb.com/airbnb-announces-partnership-with-unesco-to-promote-cultural-tourism/

Airbnb. (2021, October 7). *Delivering economic benefits to communities*. Airbnb Newsroom. Retrieved October 9, 2023, from https://news.airbnb.com/en-uk/delivering-economic-benefits-to-communities/

BBC. (2020, September 22). *France Airbnb: Paris hails victory over short-stay rents*. BBC. Retrieved October 9, 2023, from https://www.bbc.com/news/world-europe-54246005

Botsman, R., & Rogers, R. (2011). *What's Mine is Yours: How Collaborative Consumption is Changing the Way We Live*. Collins.

Cecco, L., & Willsher, K. (2019, November 19). *Airbnb faces backlash in Toronto and Paris*. The Guardian. Retrieved October 9, 2023, from https://www.theguardian.com/world/2019/nov/19/olympic-committee-deal-airbnb-angers-paris-authorities

Ganapavaram, A. (2022, November 16). *Airbnb says single-room listings jump amid cost-of-living crisis*. Reuters. Retrieved October 9, 2023, from https://www.reuters.com/business/retail-consumer/airbnb-says-single-room-listings-jump-amid-cost-of-living-crisis-2022-11-16/

Hall, C. M., Prayag, G., & Safonov, A. (2022, 09 22). *Airbnb and the sharing economy*. Taylor and Francis Online. https://www.tandfonline.com/doi/full/10.1080/13683500.2022.2122418

Inside Airbnb. (n.d.). *Data Assumptions*. Inside Airbnb. Retrieved October 9, 2023, from http://insideairbnb.com/data-assumptions

Levin, S. (2016, July 27). *Airbnb's data shows that Airbnb helps the middle class. But does it?* The Guardian. Retrieved October 9, 2023, from https://www.theguardian.com/technology/2016/jul/27/airbnb-panel-democratic-national-convention-survey

Rodríguez, D. P., Hierro, L. A., & Rodríguez-Pérez de Arenaza, D. (2019, December 30). *(PDF) Airbnb, sun-and-beach tourism and residential rental prices. The case of the coast of Andalusia (Spain)*. ResearchGate. Retrieved October 9, 2023, from https://www.researchgate.net/publication/338240066_Airbnb_sun-and-beach_tourism_and_residential_rental_prices_The_case_of_the_coast_of_Andalusia_Spain

Shih, P. (2015, November 12). *Airbnb Holds Host Convention in Paris*. Hotel News Resource. Retrieved October 9, 2023, from https://www.hotelnewsresource.com/article86514.html

Törnberg, P. (2022, 04 21). How sharing is the "sharing economy"? Evidence from 97 Airbnb markets. https://doi.org/10.1371/journal.pone.0266998

Wachsmuth, D. (2018, 02). *Airbnb and the Rent Gap: Gentrification Through the Sharing Economy*. ResearchGate. Retrieved October 9, 2023, from https://www.researchgate.net/publication/318281320_Airbnb_and_the_Rent_Gap_Gentrification_Through_the_Sharing_Economy