**Pillow Palooza**

**NYC Short-Term Rental Report**

Yarden Gamil

Date: 10/06/2023

**Introduction**

Today, Airbnb became one of the most popular services in the tourism world. Data analysis on thousands of listings provided through Airbnb data is a crucial factor for the company. This is a lot of data that can be analyzed and used for business decisions, understanding the customers' and hosts’ behaviors, guiding marketing campaigns, and more.

The data set in this report has around 25,200 observations in it, with 17 columns when merged and containing a few additional columns (from public data sources) that the team added to the data.

The main goal of this report is to provide insights into the short-term rental market in New York City. The analysis will aim to uncover key trends in popular neighborhoods, rental prices, property types, length of stay, and demand over time. The insights gained from the analysis will provide the information we need about the short-term rental market in New York City, so we could make informed business decisions and understand if this is an opportunity for us for growth and success, or whether we should give up on this project.

The process of this project includes first collecting and importing Airbnb listing data from multiple sources, cleaning the data, and then analyzing it to uncover key trends in the different dimensions of the data.

To guide me along the analysis and find the main important points for the company, I was asked to answer some questions:

1. What are the most popular neighborhoods for short-term rentals in New York City?
2. What is the average rental price for short-term rentals in New York City, and how does it vary by neighborhood and property type?
3. What are the most commonly rented property types on Airbnb in New York City, and how does this vary by neighborhood?
4. What is the average length of stay for short-term rentals in New York City, and how does this vary by neighborhood and property type?
5. How has demand for short-term rentals in New York City changed over time, and are there any seasonal trends that could impact business decisions?

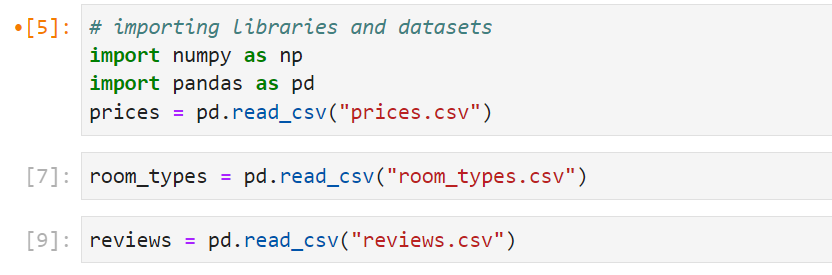
During the project, I noticed that some of the questions I can answer with the available data (Q 1-3), and for some I needed more data (Q 4-5). I will return to this subject later in this report.

**First step- Data Wrangling and Cleaning**

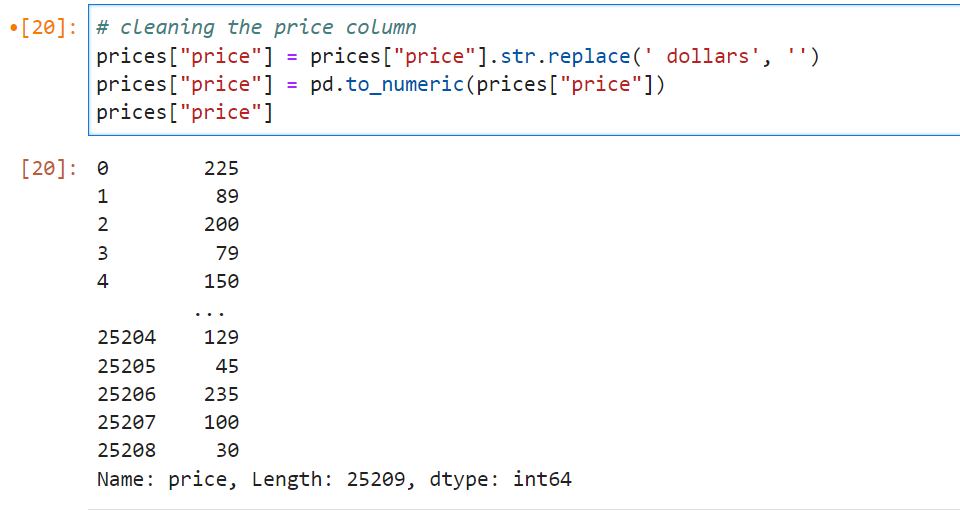
This cleaning and preparation step is critical to ensure that the data is reliable and accurate. It also sets the foundation for the next steps in the analysis, which will involve identifying trends in the NYC Airbnb market.

**Importing, cleaning, and preparing the NYC Airbnb dataset using Python  
(for the following analysis in SQL) :**

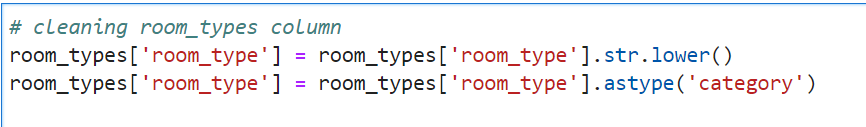
As a first step, using JupiterLite I imported the three relevant datasets we have for this project, explored them, and did a cleaning process. This step was done in Python. To import the datasets I used the following code in the notebook:



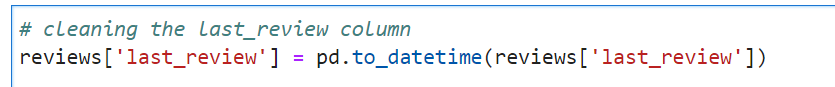
Looking at the datasets, I noticed that the price column in the prices DataFrame currently records each value as a string with a currency (dollars) following, i.e., “225 dollars”. To clean the column I removed whitespace and string characters trailing the numeric values, then updated the "price" column to a numeric dtype.



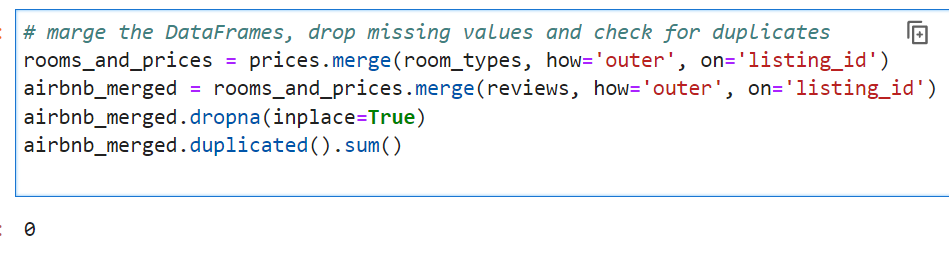
The room\_type column needed to be cleaned as well. We are interested to divide the column into three types of rooms (entire home/apt, private room, and shared room). I noticed that there are several variations for the listings. To make them uniform I first converted all string characters to lowercase, and converted the column to a ‘category’ type. I will need this later to be able to count the frequency of each room type.



The last\_review column contains the date of the last review in the format of "Month Day Year" e.g., May 21 2019. I needed to ensure that the format is right and the same along the column, to be able to keep the analysis later. For this purpose, I converted all the dates to “date\_time” format with the use of the pd.datetime() method.



From here I continued to analyze the three DataFrames, but it is also worth mentioning that during the analysis I also merged them to make any future analysis easier to conduct. Due to the large number of columns present, I decided to check a few of them first, in the process of cleaning the relevant and needed columns, and not end up cleaning all the columns in the data at once. Once I have joined the data (I called it "airbnb\_merged”), I removed any observations with missing values and checked for duplicates.

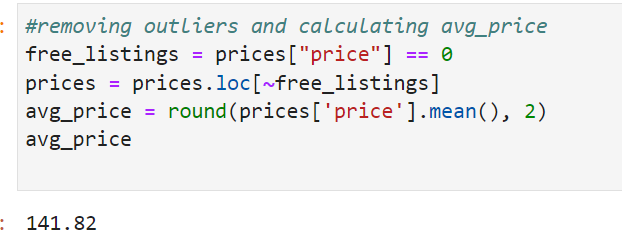


Again, this was done in the later stage of the analysis here in Python, but we can see that no duplicate values were found here.

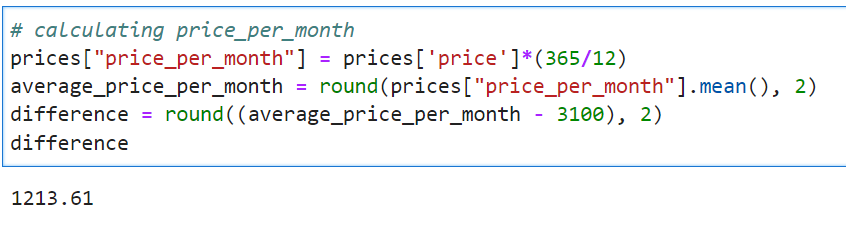
From here I moved to explore the DataFrames in the notebook. First I used the describe() method on the prices table.



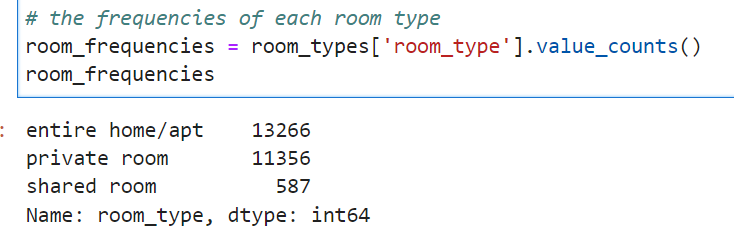
we can see here that three quarters of listings cost $175 per night or less. However, there are some outliers including a maximum price of $7,500 per night. Some listings are even showing as free. To fix this I removed the outliers from the DataFrame and calculate the average price instead:



I got how much a listing costs, on average, per night, but it will be useful to have a benchmark for comparison. With the fact that 1 bedroom apartment in NYC costs, on average, $3,100 per month, I convert the per night prices of our listings into monthly costs, so I could compare it to the private market.

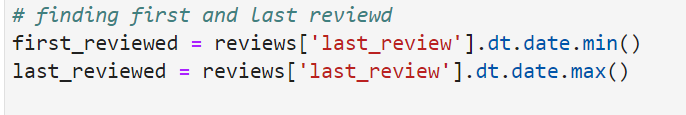


From here I could see that using Airbnb appears to be substantially more expensive than the private rental market, but because these Airbnb listings include not only single private rooms, I also counted the values for each room\_type:

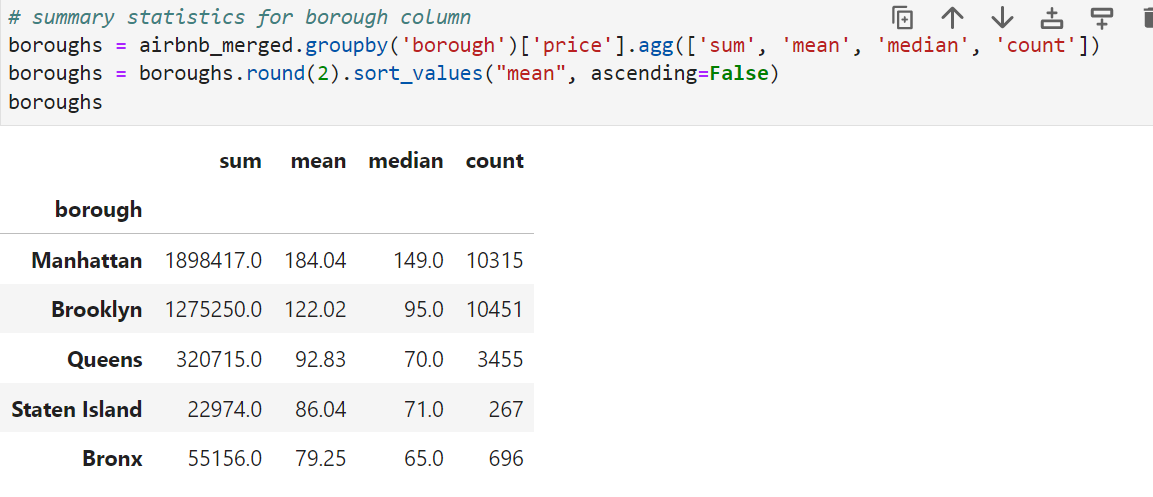


It seems that there is a fairly similar-sized market opportunity for both private rooms (45% of listings) and entire homes/apartments (52%) on the Airbnb platform in NYC.

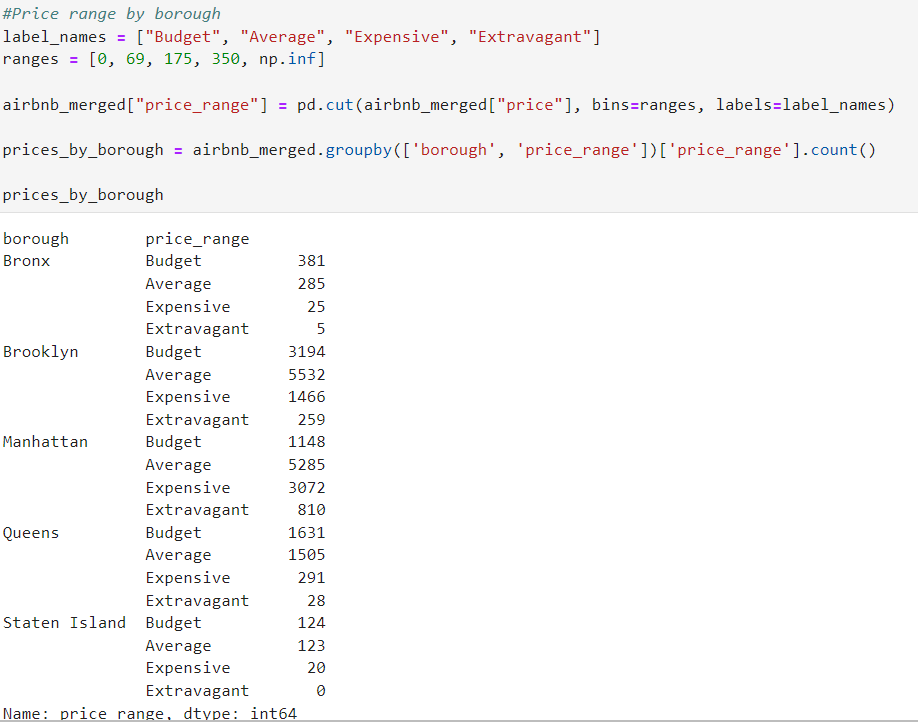
Moving to the reviews DataFrame, and after cleaning earlier the dates to the right format, I found out the earliest and latest ‘reviewed’ dates in the DataFrame, using the “.dt.date “ method. It was found that the first review was on 2019/01/01, and the last review on 2019/12/06.



After I combined all data into a single DataFrames as I explained earlier in this report, I also put attention to the difference in listing prices between New York City boroughs. For this purpose, I calculated summary statistics using the groupby() method.



The above output gives us a summary of prices for listings across the 5 boroughs. I wanted to categorize the listings based on whether they fall into specific price ranges, and view this by borough. It was done by using percentiles and labels to create a new column- price\_range, in the DataFrame, then grouped the data and counted frequencies for listings in each price range by borough.



It looks for example that there is not even one listing in Staten Island which is extravagant, and in Brooklin and Manhattan we can find mostly rooms at average prices (according to the labels we set).

**Second step- Analysis and Results**

Now that the data is clean and I started to explore it a bit, I could dive into the main analysis process to find important insights, with also considering the questions I was asked to answer in the meeting. For this stage I moved to use SQL, after exporting the clean and preprocessed data, and with additional columns (from public data sources) that the team added to the data, including the schema for this analysis. The schema and the entire analysis in SQL are in the appendix (queries and results). The main insights from this analysis are:

**\*** **The most common room type** in the entire NYC Airbnb listings is Entire home/Apt.

Following this, the average price for this room\_type is about 197 dollars. It looks like for people or families that want to rent an entire home/apt- a price around this amount looks reasonable and manageable. Also, we should consider focusing more on this room\_type in our promotions.

**\*** following the last point, it is reasonable that most of the listings with a price over 500$ per night are found in this room type because it’s the largest out of the 3 types.

**\*** **looking into boroughs,** most entire home/apt listings are in Manhattan and Brooklyn. Those boroughs also have the most listings overall (in any room\_type). It should be considered that those areas are more popular among tourists and that the range of prices in them is also the widest- where Manhattan's average price per night starts at 10$ and reach up to 5,100$, while for Brooklin the average price range is from 10$ and up to 7,500$. Due to the vast range (from a very cheap to a very expensive price), those boroughs may be useful for targeting a variety of types of guests. Also, it will be worth putting less effort into Staten Island, where only a few hundred listings are there. It’s worth noting that those findings are also logical because it’s known that Manhattan is one of the most expensive places in the world to live in, where Staten Island or Bronx for example are on other hand appears to have lower standards of living.

**\*** **The most profitable borough** in terms of revenue for hosts is again Manhattan, with about 395,500,000$, followed by Brooklyn, with a little above 279,000,000$. This is another strong evidence of the popularity of those two boroughs.

**\*** in an entire home/apt room type, Sea Gate neighborhoodhas the **highest price on average per month** for all the listings there. For a shared room, Vinegar Hill is the the highest, and for a private room- West Village. **There are no listings without reviews in the data**, which is a good point to know and make an effort to keep this trend.

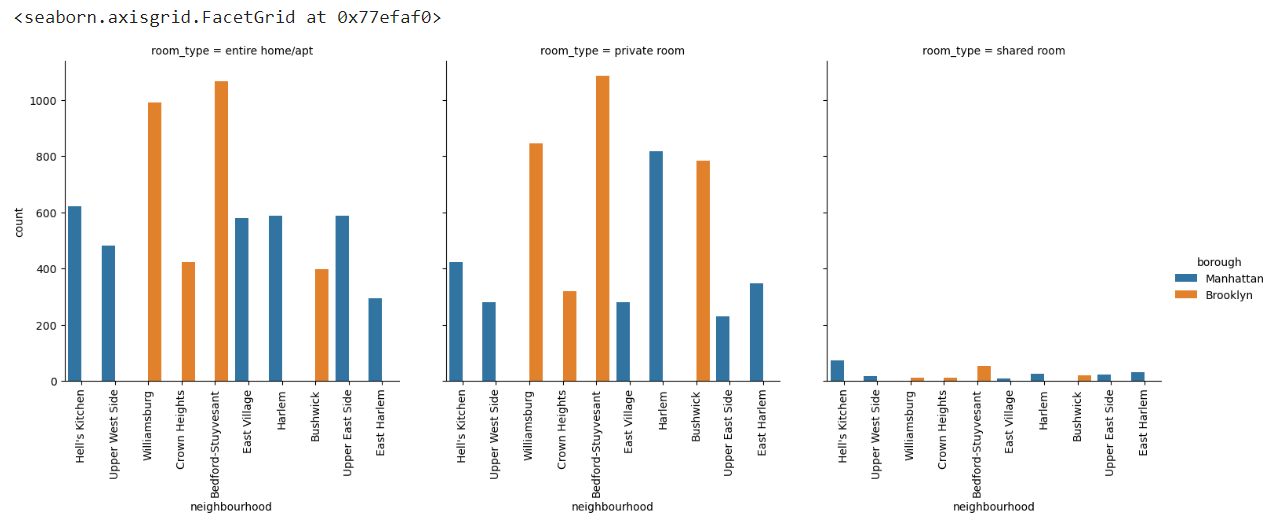
**\*** When checking **the correlation between estimated book days with the prices** of Airbnb listings in New York City, there is a very weak (below 0.1) negative correlation between the two parameters. This means that for the estimated booked days there is no meaning when looking at the prices and vice versa, but if already there is a very weak correlation, the two are going in opposite ways (i.e. less days-higher price. But again, the correlation is very weak to almost not appear).

I thought it will also be valuable to look at things visually. I was interested to see which 10 neighborhoods have the most listings in the combined data (**Top 10 neighborhoods**), and how is **their relativity/variability between the different room types**. To make this visualization I returned to the notebook and used Python to import the relevant modules to be able to receive the desired output. For this plot I used seaborn catborn visualization.

In the following viz, you can see the three subplots (one for each room type, and the color refers to the name of the borough, so each borough gets its own color-which is the same in all the subplots). Each subplot represents the top 10 neighborhoods we found on the x-axis, and their count in each room\_type on the Y-axis.







This plot shows us very clearly a few interesting things. First, the shared rooms are very not popular compared to the other room types among the 10 most populated neighborhoods (by number of listings). The difference is so big that I recommend giving up on this category of room in those neighborhoods and focusing there only on entire homes/apt, and private rooms.

Second, we can see here that only 2 boroughs are represented in the top 10 neighborhoods- Manhattan and Brooklyn- two boroughs that were found popular earlier in the analysis. We can also notice that Harlem and Hell's Kitchen are the most popular for Manhattan borough, and Williamsburg and Bedford-Stuyvesant for Brooklyn (along the different room types).

**Conclusion and Recommendations**

This dataset appeared to be a very rich dataset with a variety of columns that allowed us to do a deep data exploration with the aim to find meaningful insight and make the best decisions for the company to succeed.

First, we found that the price per night of all the listings in the data has a very wide range, with some extreme values. Later analysis showed us that those were due to a few areas like Manhattan and Brooklin boroughs, which are the most expensive in this term. We could see it by categorizing the listings based on whether they fall into specific price ranges, and also by visualization, which added the neighborhood and room type aspects.

After that we used the price (per night) column to calculate the monthly price for each listing, to be able to compare it with the private rental market. It was found that using Airbnb is more expensive than the private market, but considering the different room types we saw that most of the listings are either entire rooms/apt or private rooms, which is an important mark when we think about what the hosts and tourist preferences, and what room types should we focus on.

Then we moved from Python and used SQL to query the merged dataset and deep more to find interesting insights. We found that the average revenue was the highest for Manhattan and Brooklin, which is consistent with our last findings.

When looking into neighborhoods and room types, we saw that Sea Gate neighborhood has the most highest price on average per month for an entire home/apt, for a shared room it is Vinegar Hill which is highest, and for a private room- West Village.

Those neighborhoods could be targeted, considering the most popular room type in each of them.

In the last query in SQL, we searched for the correlation between the estimated book days with the price of an Airbnb listing in New York City, but we couldn’t find a strong correlation between those features, to be able to make smart decisions by considering those two together.

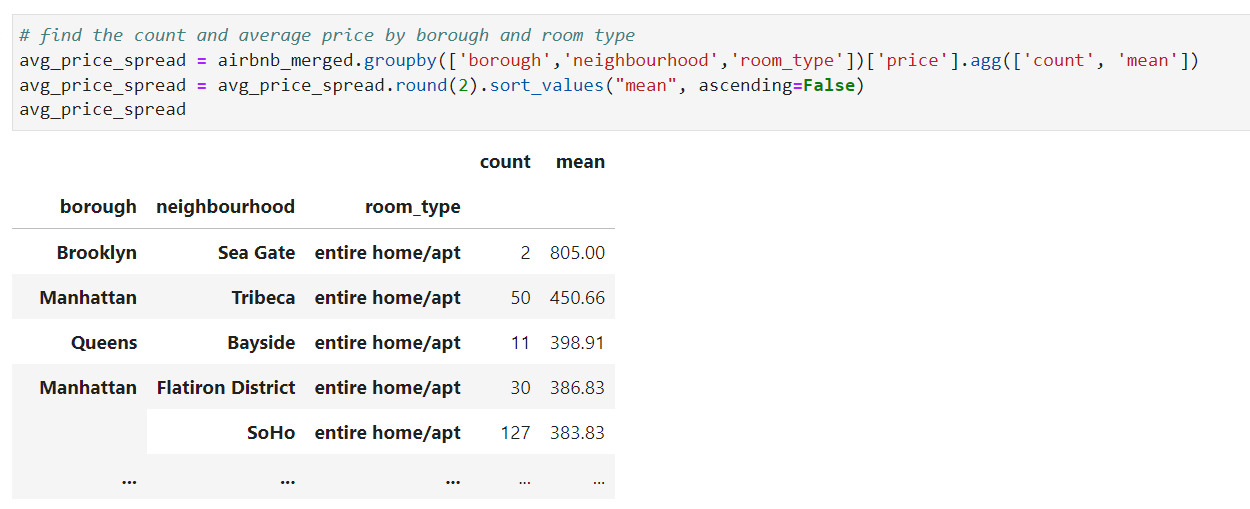
Back again to the 5 questions I got during the meeting as described in the introduction, some of them were able to be answered with our available data (questions 1-3). Let's deep into them:

1. **What are the most popular neighborhoods for short-term rentals in New York City?**

As we saw, those are the top 10: Bedford-Stuyvesant, Williamsburg, Harlem, Bushwick, Hell's Kitchen, East Village, Upper East Side, Upper West Side, Crown Heights, and East Harlem.

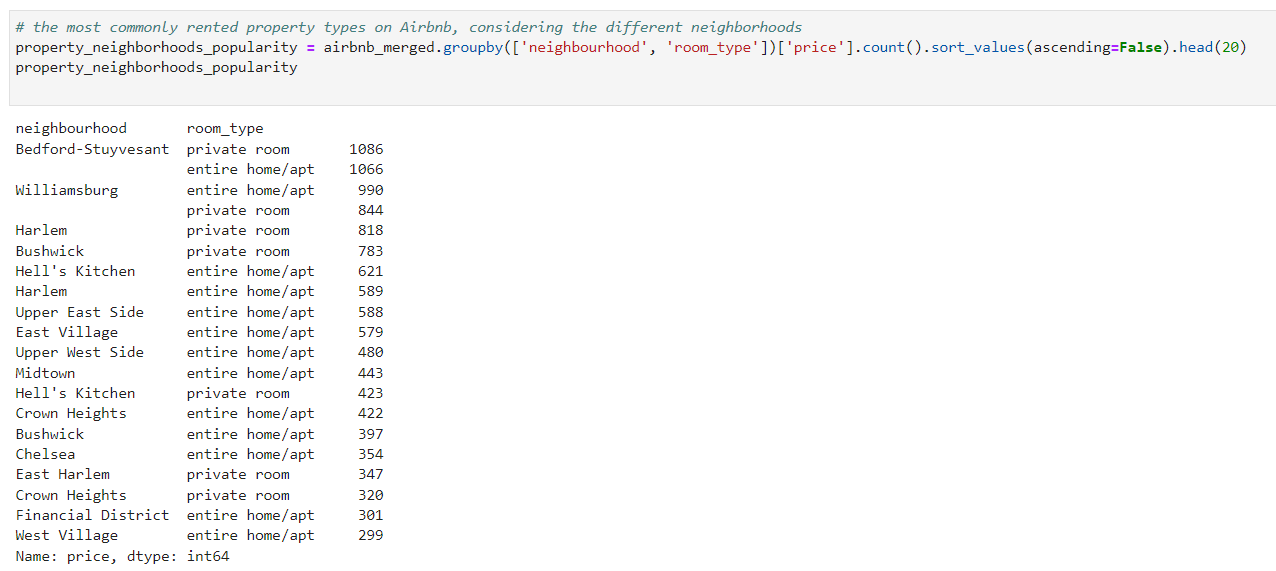
1. **What is the average rental price for short-term rentals in New York City, and how does it vary by neighborhood and property type?**

With the code down we can see that Sea Gate and Tribeca, when the room type is ‘Entire home/Apt’- are at the top with the highest average price along the DataFrame. Furthermore, those two neighborhoods belong to the most popular boroughs in the data- Brooklyn and Manhattan. From the count and average price, we see clearly that those specific combinations **aren’t** a big part of the entire data, so it may be worth promoting those specific areas to the customer segment that has the ability to stand in those prices.



1. **What are the most commonly rented property types on Airbnb in New York City, and how does this vary by neighborhood?**

To find the answer I used the following code:



We can see that the most commonly rented properties are private rooms, and entire homes/apt, which is supported by previous findings, but it's interesting to see that in each neighborhood the common property between those two is different. When looking at the most rented neighborhoods we could see that in Bedford-Stuyvesant for example, those two room types are common pretty evenly. In Bushwick on the other side, private rooms are rented twice as more as entire homes/apt. those findings (and about other neighborhoods) are crucial when we try to target the right customers for each place, and also for promotional campaigns.

When talking about the 2 last questions I was asked, I found that the data we have here isn’t enough to answer them.

For **question 4- What is the average length of stay for short-term rentals in New York City, and how does this vary by neighborhood and property type?**

The provided dataset doesn’t include data on the length of stay, which means we have partial information to answer this. However, it will be beneficial to have this information, to be able to go even deeper into the data we already have.

following **Question 5- How has demand for short-term rentals in New York City changed over time, and are there any seasonal trends that could impact business decisions?**

Also here, The provided dataset doesn’t allow us to answer this question with full certainty. To answer this question we need to also have the information about the date of each booking-preferably in another table that includes the information, while every record (row) represents one booking and has its own unique id. Thus, we could combine this table with our merged dataset to be able to see the changes over time. Nevertheless, with the existing data, we can see the changes in short-term rental in NYC over different boroughs or different neighborhoods.

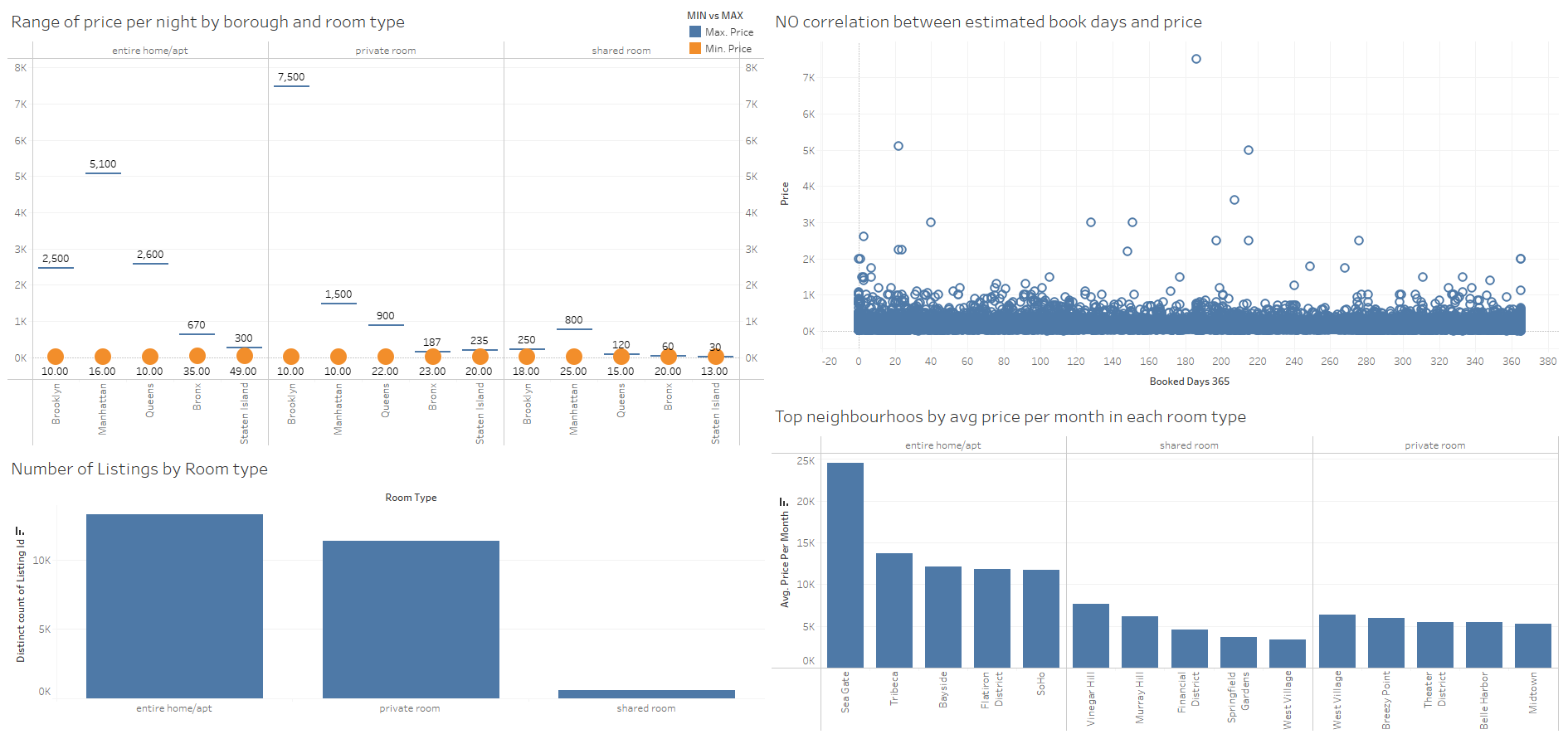
For our data exploration purposes, it also would be nice to have a couple of additional features, such as the average review score for each listing (on a scale of 1-5, or 1-10), so we could look for the best-reviewed hosts in NYC.

Overall, we discovered a very good number of interesting relationships between the features and explained each step of the process. From all that has been said I think that although it will be nice to have more information, we should launch the project using all of the insights we gained, and keep looking after more new features and records to expand our analysis and knowledge about the NYC rental market.

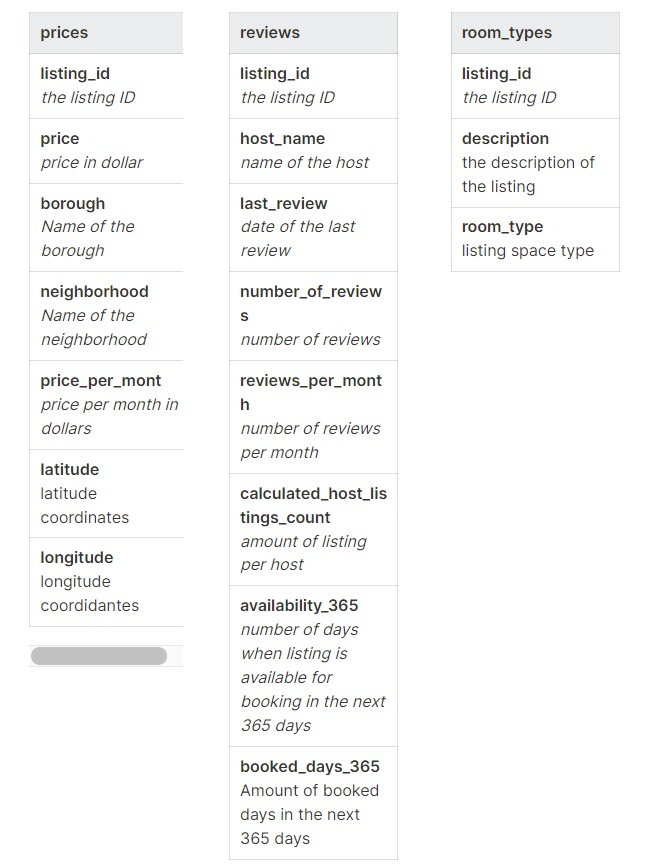
**Appendix**

**Tableau:**

<https://public.tableau.com/views/NYCshort-termrentalanalysis/Dashboard1?:language=en-US&:display_count=n&:origin=viz_share_link>



**Schema table:**



**SQL analysis:**

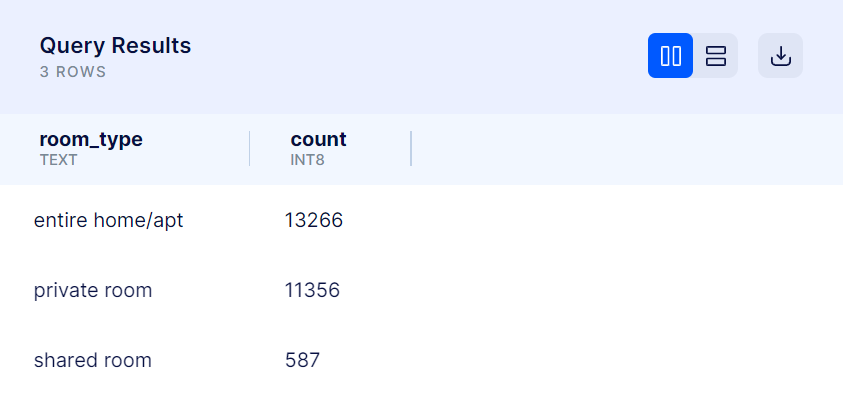
**1. What is the most common room type in NYC Airbnb listings?**

SELECT room\_type, COUNT(room\_type)

FROM room\_types

GROUP BY room\_type

order by 2 DESC;



**2.** **What is the average price of a listing by room type?**

SELECT room\_typE,

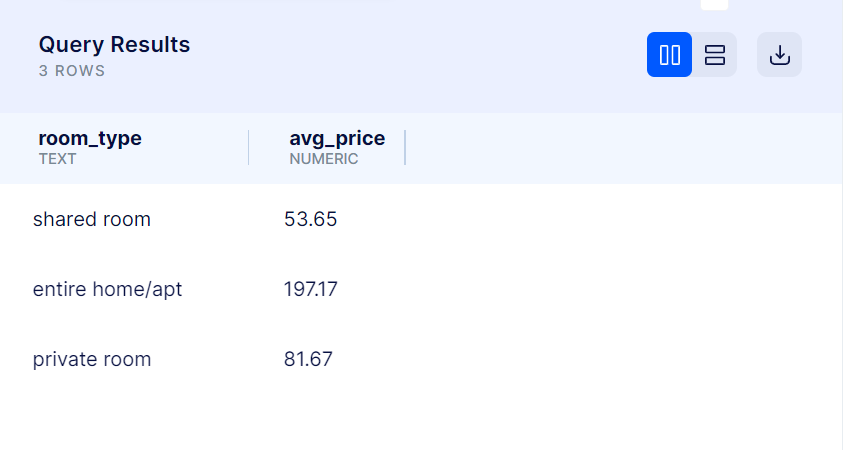
    ROUND(AVG(price::NUMERIC), 2) as avg\_price

FROM room\_types rt

JOIN prices p

ON rt.listing\_id = p.listing\_id

GROUP BY room\_type;



**3. Which borough has the highest average price per month?**

SELECT borough,

    ROUND(AVG(price\_per\_month::NUMERIC), 2) as avg\_price\_per\_month

FROM room\_types rt

JOIN prices p

ON rt.listing\_id = p.listing\_id

GROUP BY borough

ORDER BY 2;



**4. How many listings of each room type are in each borough?**

SELECT borough, room\_type,

    COUNT(room\_type) as number\_of\_listings

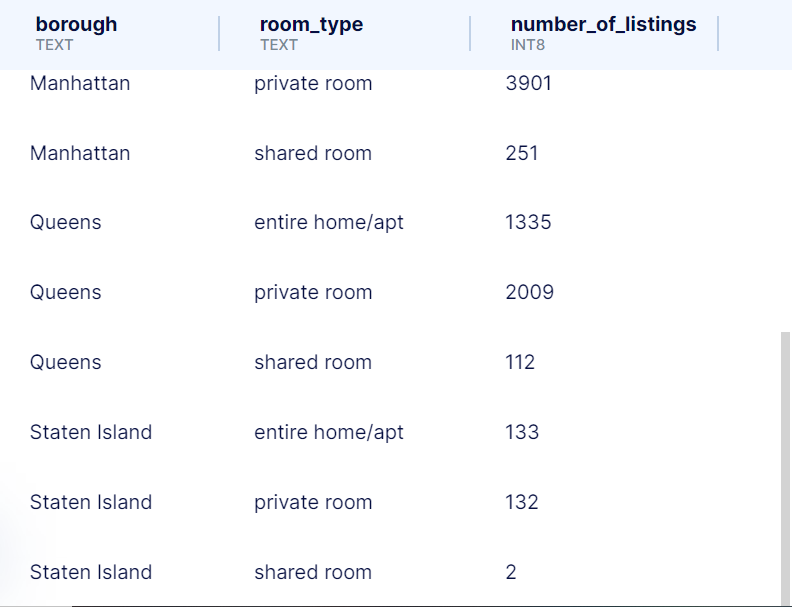
FROM room\_types rt

JOIN prices p

ON rt.listing\_id = p.listing\_id

GROUP BY borough, room\_type

ORDER BY borough, room\_type;



**5. How many listings in each room type category have a price of over $500 per night?**

SELECT room\_type,

    COUNT(p.listing\_id) as listings\_over\_500\_pn

FROM room\_types rt

JOIN prices p

ON rt.listing\_id = p.listing\_id

WHERE price > 500

GROUP BY room\_type

ORDER BY listings\_over\_500\_pn;



**6. What is the distribution of listing prices by borough?**

SELECT borough,

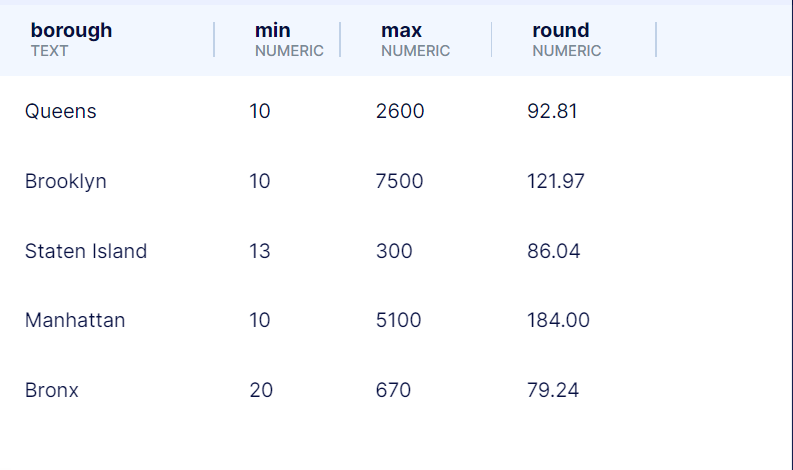
    MIN(price)::NUMERIC, MAX(price)::NUMERIC, ROUND(AVG(price)::NUMERIC, 2)

FROM room\_types rt

JOIN prices p

ON rt.listing\_id = p.listing\_id

GROUP BY borough;



**7. What is the estimated amount of revenue generated by hosts in each borough?**

SELECT borough,

    ROUND(SUM(price \* booked\_days\_365), 2)

FROM room\_types rt

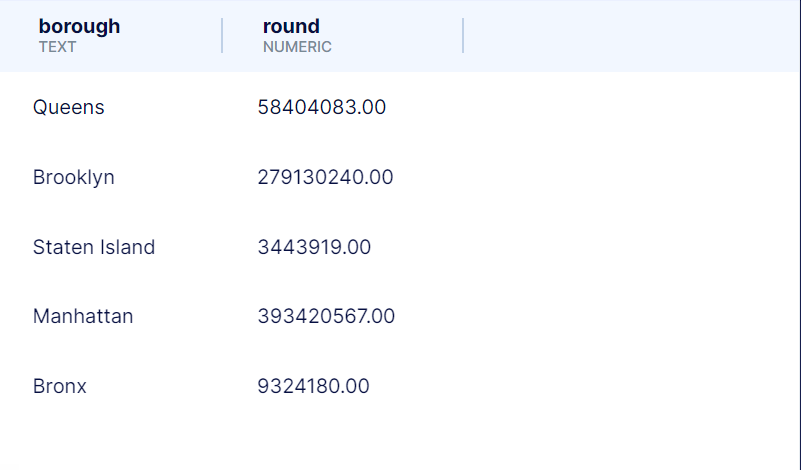
JOIN prices p

ON rt.listing\_id = p.listing\_id

JOIN reviews re

ON p.listing\_id = re.listing\_id

GROUP BY borough;



**8. What is the average price per month for listings in each neighborhood?**

SELECT neighbourhood, room\_type,

    ROUND(AVG(price\_per\_month::NUMERIC), 3) as avg\_price\_per\_m

FROM prices p

JOIN reviews re

ON p.listing\_id = re.listing\_id

JOIN room\_types rt

ON p.listing\_id = rt.listing\_id

GROUP BY neighbourhood, room\_type

ORDER BY room\_type, avg\_price\_per\_m desc;



**9. How many listings have no reviews?**

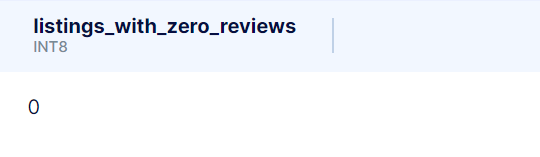
SELECT COUNT(p. listing\_id) as listings\_with\_zero\_reviews

FROM prices p

WHERE p.listing\_id NOT IN (

    SELECT re.listing\_id from reviews re

)



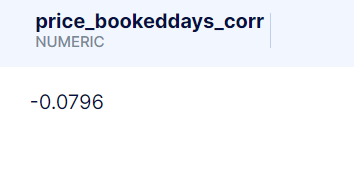
**10. How do the estimated book days correlate with the price of an Airbnb listing in New York City?**

SELECT ROUND(corr(p.price, re.booked\_days\_365)::NUMERIC, 4) as price\_BookedDays\_corr

FROM reviews re

JOIN prices p

ON re.listing\_id = p.listing\_id

****