New York Airbnb Dataset Analysis

The reason I choose this topic is that I will visit USA this summer. Airbnb will be a necessity. Taking a deep look into the dataset may help me have more understanding on the accommodation system. Then maybe I can make a decision more wisely.

In this case, we are given three questions from the instructor. The detail process of solutions is written in the main.ipynb file. Here I list the answers to the questions briefly and comment on the results.

Question 1:

In the data, there are two values of “host_identity_verified”. Which value is larger?

Answer:

Unconfirmed: 50944

Verified: 50825

Number of the ‘unconfirmed’ value is larger.

Question 2:

What are the top 2 “neighbourhood_group”?

Answer:

Average neighborhood group rate number:

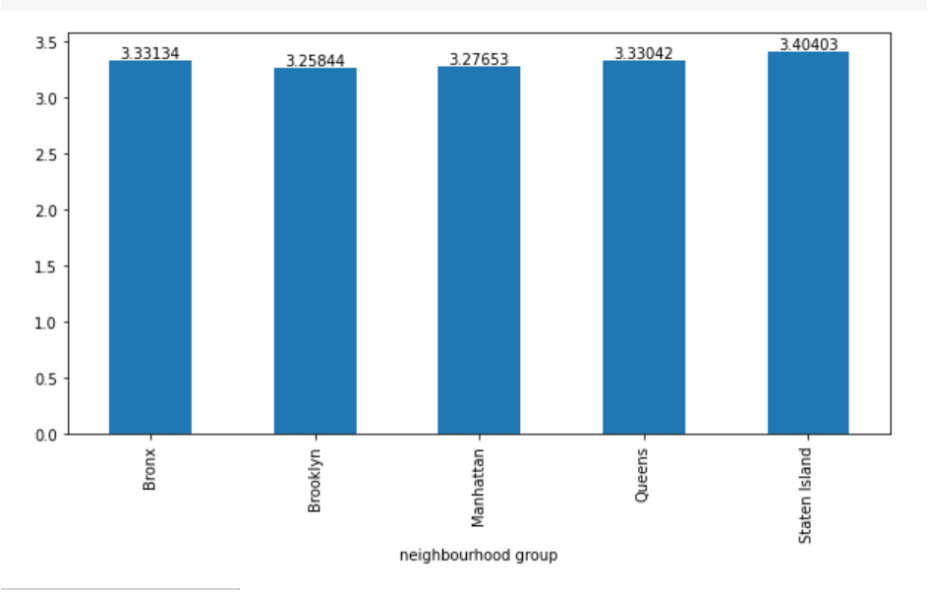
Staten Island: 3.404030

Bronx: 3.331341

Queens: 3.330420

Manhattan: 3.276528

Brooklyn: 3.258443



As the result, the top two neighborhood group are Staten Island and Bronx.

Question 3:

How many room types are in the data and what are their proportions?

Answer:

There are four types of room in the dataset.

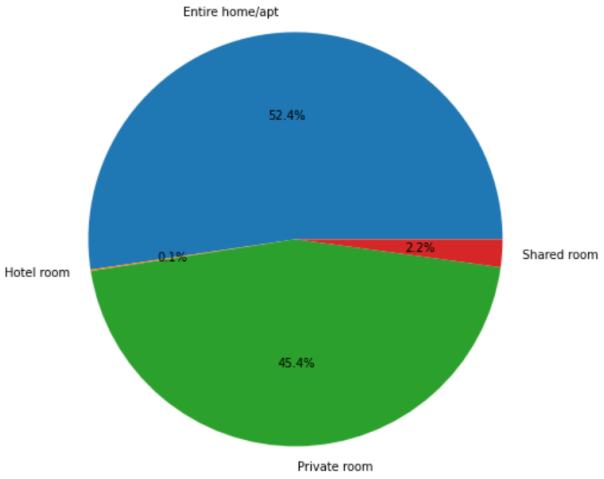
The proportion:

Entire home/apt: 52.4%

Hotel room: 0.1%

Private room: 45.4%

Shared room: 2.2%

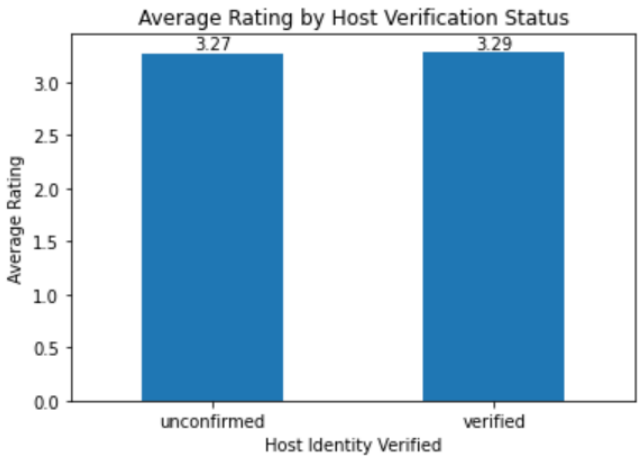


Additional questions:

Question:

Does choosing a verified host impact my experience? What features make a host more likely to receive high ratings?

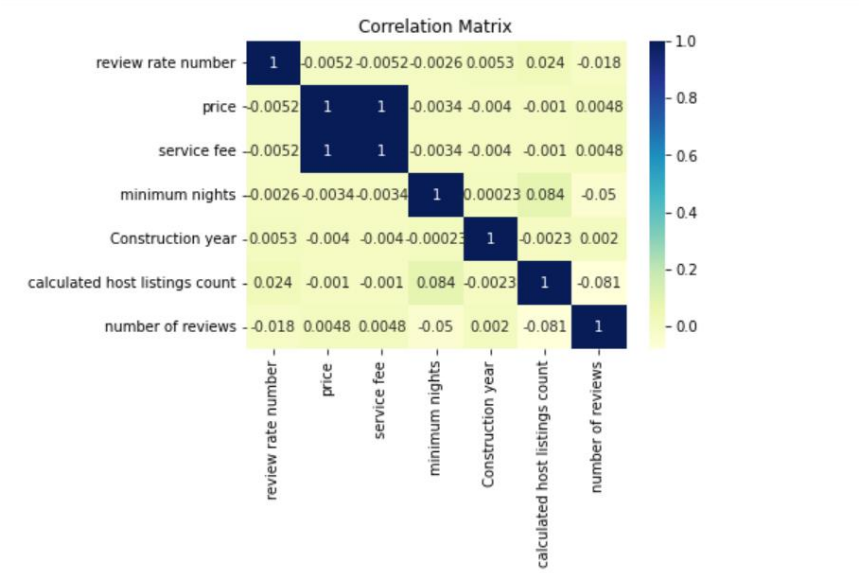
Answer:



From this result, the verified ones are rated higher than the unconfirmed ones. But they are very close just like q2. We can't actually tell there is a strong impact on the verified status.

Thus, I choose to calculate the correlation between rate and interesting features that are linear to see the relationship.

Result:

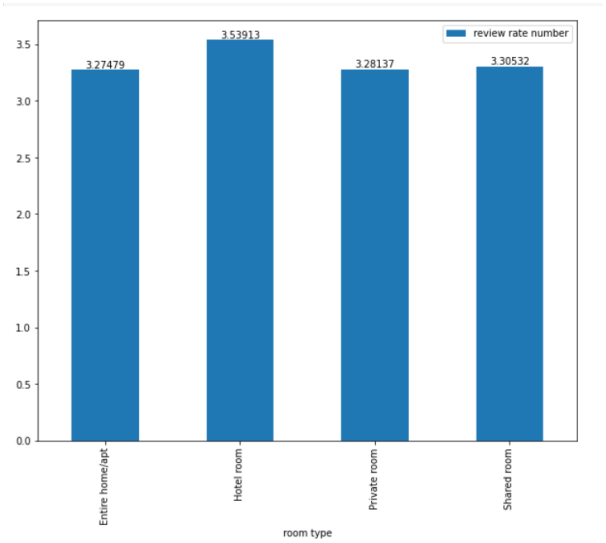


Sadly, the correlations are all weak. We can barely have conclusion based on the result. But I can still find some observations:

The one between calculated host listings count and review rate number is also positive and moderate (0.024), suggesting a moderate positive linear relationship between these two variables. This may indicate that the hosts with more listings are also more experienced on setting a Airbnb then get higher rate of review.

I also want know which room type is rated higher in the dataset.

Result:

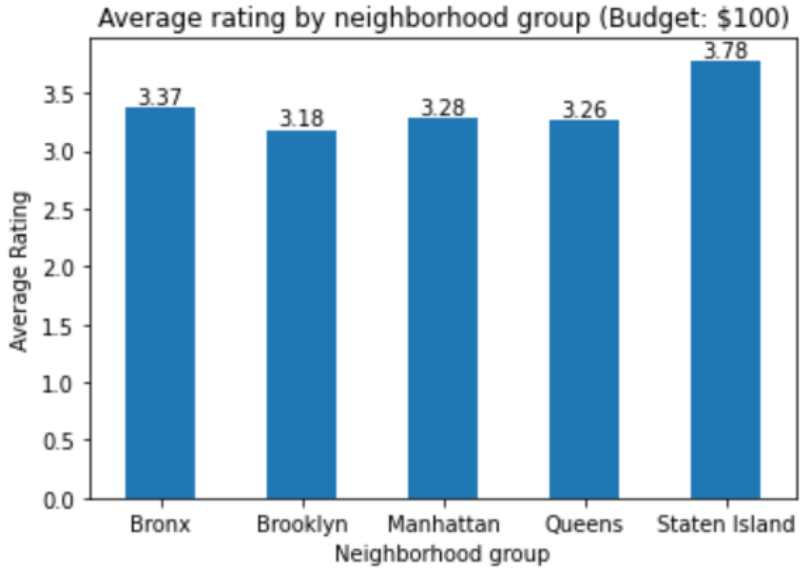


The ‘Hotel room’ is rated highest which is not a surprise. However, the price is also an issue. I can’t stay at a hotel room every day. I need to take price of each room type into consideration.

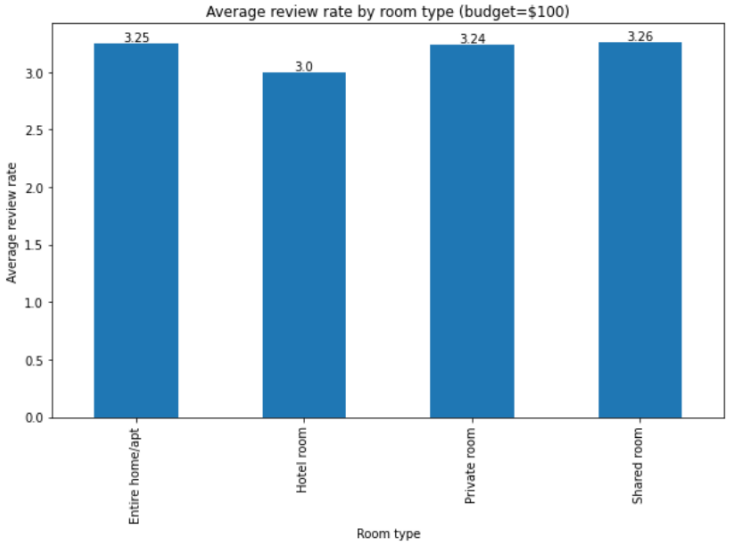
Question:

Because my budget is limited, I wonder which neighborhood group is rated highest, and the price is lower than the limitation (budget = $100).

Answer:



Based on the results, it seems that the neighborhood group with the highest average rating within the budget of $100 is Staten Island with a rating of 3.78. However, it's worth noting that the number of listings in Staten Island(46) might be much smaller than other groups like Manhattan(1938) and Brooklyn(1831), which could affect the accuracy of the average rating. Let’s move on to the room type.



Under the budget, staying at a hotel room is no longer a good idea. The shared room type seems to be a better option.

Other Findings:

1. There is a weak negative correlation between the price and the review rate, indicating that lower-priced listings tend to receive slightly higher review rates.
2. The majority of the listings are concentrated in Manhattan and Brooklyn, with over 90% of the listings located in these two boroughs.
3. Private rooms and Shared rooms have a slightly higher average review rate than Entire home/apts overall.
4. The number of reviews is negatively correlated with the price, indicating that lower-priced listings tend to receive more reviews.
5. The year of construction of the building does not seem to have a significant impact on the price or review rate.

Overall, although the dataset only provides information in NYC, I can still use it to make more informed decisions when choosing an Airbnb, such as realizing what factors should I take into consideration and which features may have more impact on the review rate. Nevertheless, some observations I tried to find ended up ambiguous or irrelevant results. I hope to continue improving and learning data analysis skills, and in the future, be able to generate even more value from a dataset.