Project Report: BnB Analysis

Andres Perez

TABLE OF CONTENTS

Task 1

Data Overview

Our initial step was analyzing the Torerobnb_listing_rev.csv dataset where we evaluated the quality and cleanliness of the data. We recognized some inconsistencies on the data that would make it hard to create accurate and insightful data analysis. Hence, we performed a data cleaning process. This first initial step is key because later on it will help us ensure reliable results and meaningful data interpretations. Below are the data cleaning steps:

1. Column Elimination:

Redundant Identifier Removal: We removed the ID column as it does not contribute to our analysis of rental trends and profitability.

2. Data Point Exclusion:

Price Anomalies: We observed that rows 2 to 31 in the ToreroBNB dataset listed the prices as zero. Recognizing that data imputation techniques would not accurately estimate nightly prices for these listings, and considering the dataset's ample size, we decided to exclude these rows. Apart from these rows, we also eliminated all rows that did not have a price listed. This decision is based on the rationale that their absence would not significantly impact the overall predictive quality of our analysis.

3. Column Transformation and Creation:

Bathroom Data - We split the original Bathrooms column into two new columns for improved clarity and usability:

- Bathrooms: This new column isolates the numerical value indicating the quantity of bathrooms in a listing.
- Shared: This column categorizes the type of bathroom (e.g., shared Vs. not shared represented as a 0 or 1).

4. Date Conversion:

Standardizing Date Information: The Last_review and Host_Since columns were transformed to reflect only the year, converting these dates into numerical values. This simplification is meant to help when doing temporal analysis and trend identification.

5. Variable Conversion:

Categorical to Factor Transformation: To facilitate our analysis, we converted several categorical variables into factor variables. This includes:

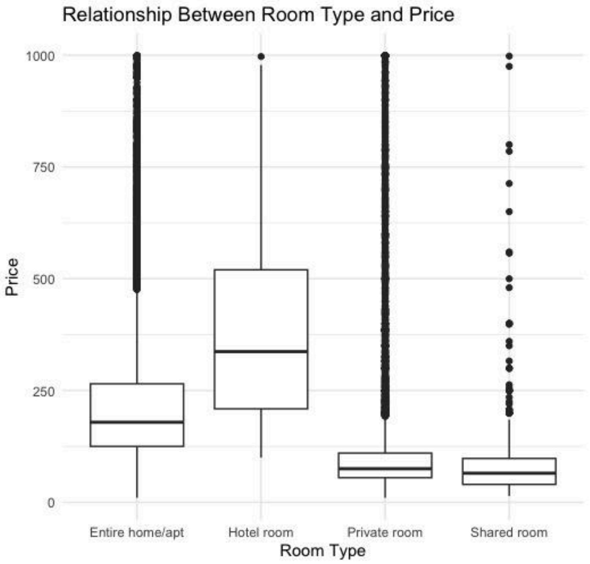
- Neighborhood
- Neighborhood_group
- Room_type

The cleaned dataset is now ready for in-depth analysis.

Data Visualization & Analysis

In this section we use various plots to visualize key aspects of the data and understand relationships between variables, especially those we believe have an effect on the price.

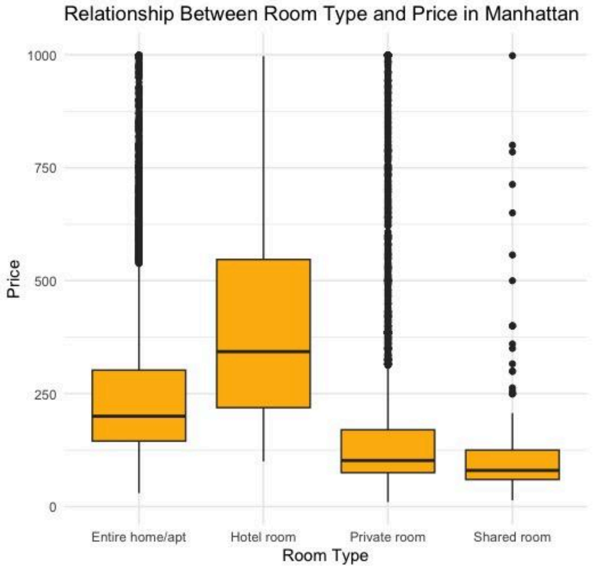
1. **Exploring relationship between room type and price:** We believed that the kind of room might be a major factor in setting the price. To see if this is true, we created scatterplots to show the possible connection between room type and price.



This first boxplot shows the relationship between room type and price. As expected, hotel rooms on average are the most expensive out of all categories. This may be due to the services included with being in the hotel and most hotel rooms have a private bathroom. The next most expensive type is having an entire home/apartment, then renting a private room and finally the cheapest option is sharing a room.

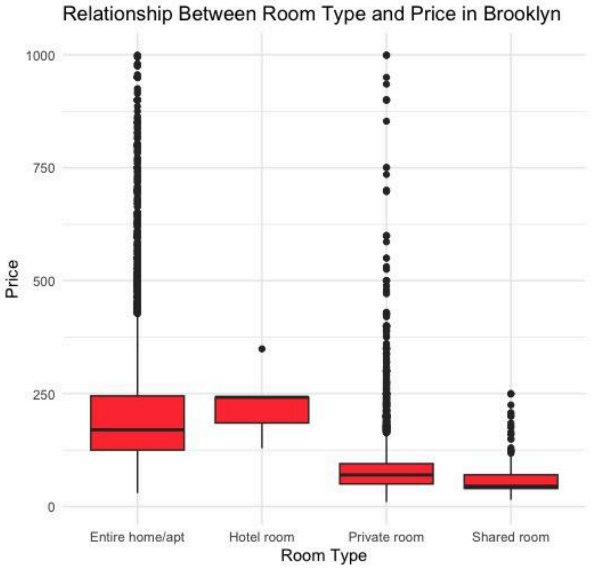
We now evaluate this relationship per borough.

Manhattan:



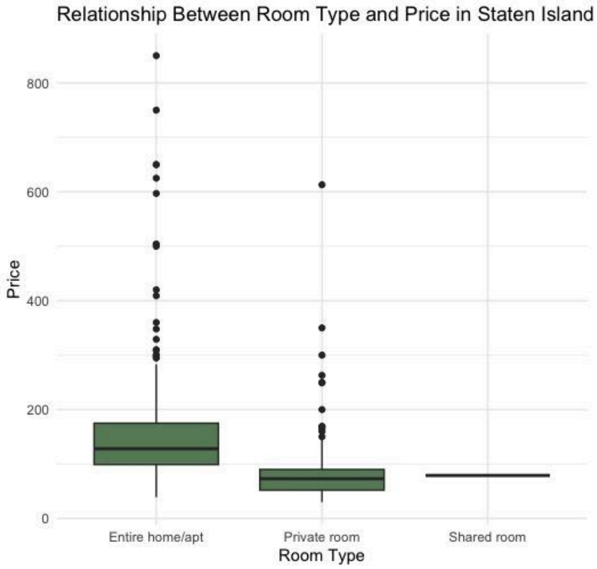
Relationship Between Room Type and Price in Manhattan

Brooklyn:



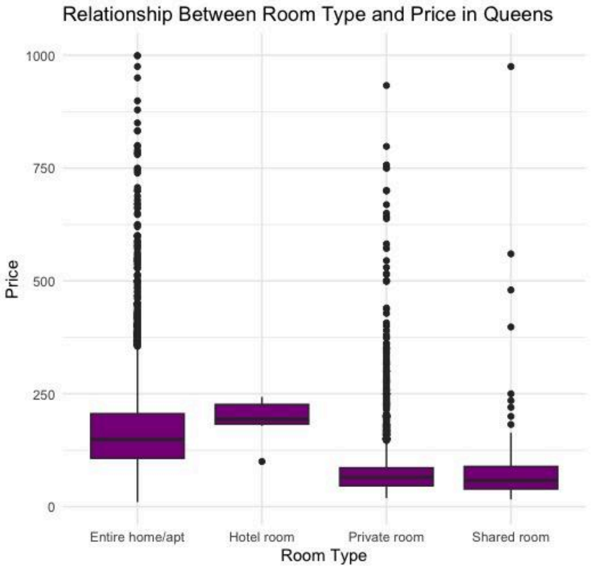
Relationship Between Room Type and Price in Brooklyn

Staten Island:



Relationship Between Room Type and Price in Staten Island

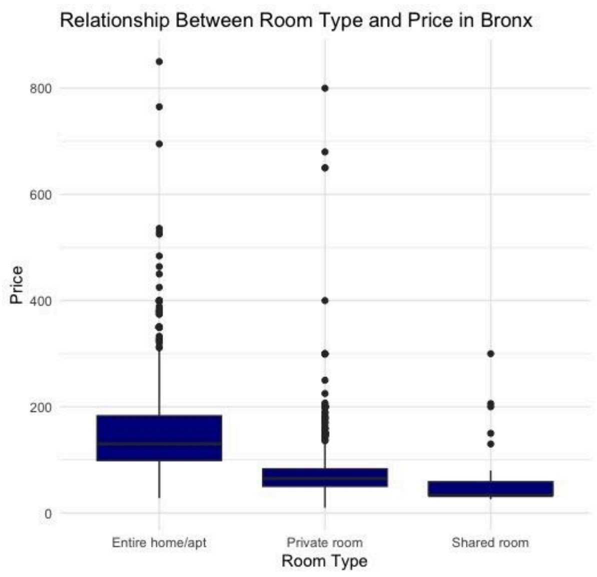
Queens:



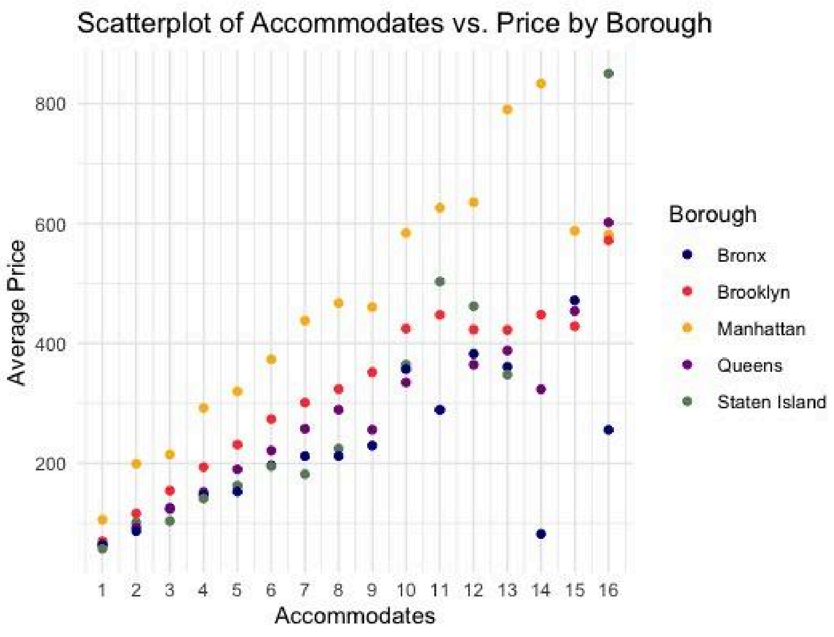
Relationship Between Room Type and Price in Queens

Bronx:

Conclusion: Looking at the five boroughs, it's interesting to see that while total prices vary from one borough to another, the order of room types by cost stays the same - Hotel rooms are the most expensive and

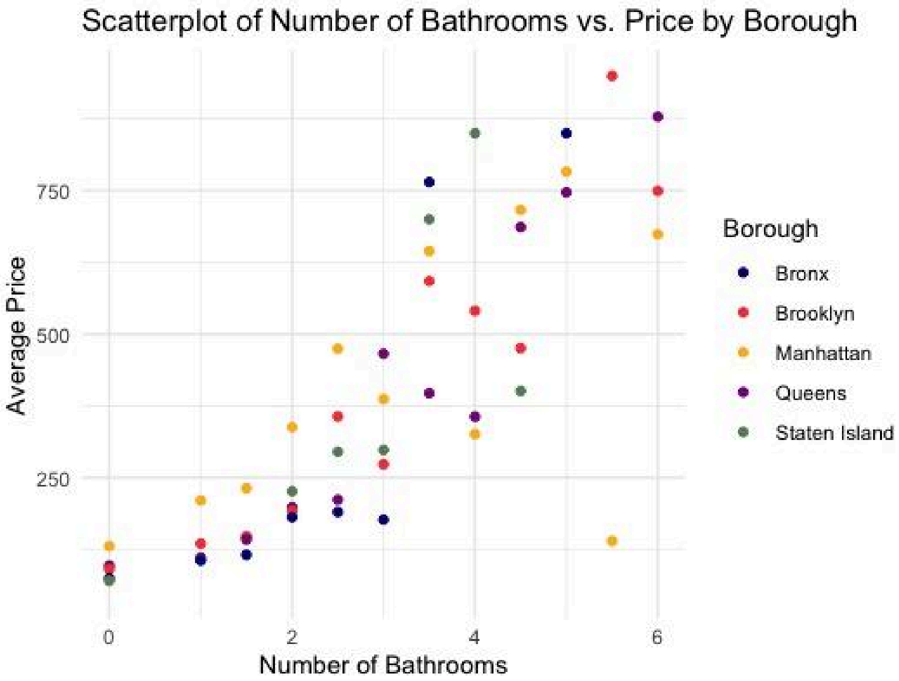
	shared rooms are the least. We also observed that in Staten Island and the Bronx, the two least expensive boroughs, there's no information about hotels. This could be for two main reasons. Firstly, these areas might not attract many tourists. Secondly, factors like safety or lack of cultural activities might make these boroughs less appealing, so hotels might not be interested in establishing there.

2. **Exploring relationship between accommodates and price**



This scatterplot clearly shows the link between the listing's maximum capacity (Accommodate) and its price. Across all five boroughs, this relationship appears quite linear. As the capacity for guests in a house or rental property increases, so does the price. This is logical, as larger places with more space generally cost more.

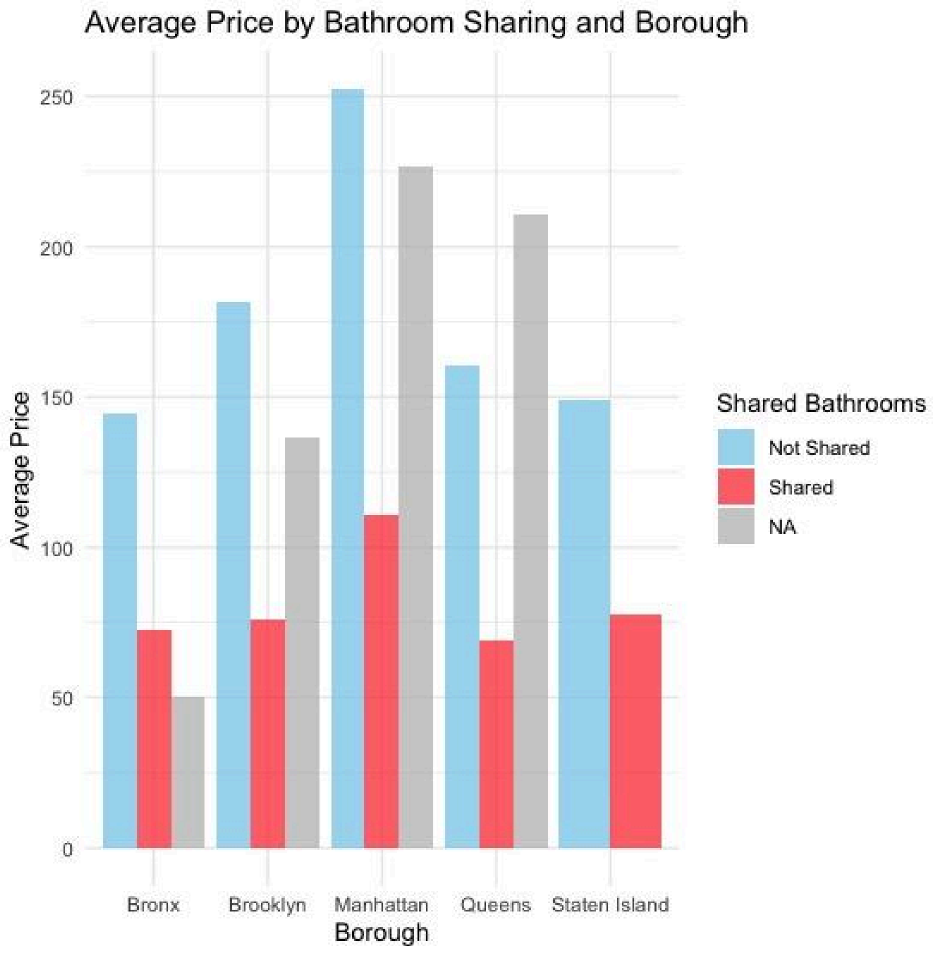
3. **Exploring the relationship between bathroom and price**



The connection between the number of bathrooms in a property and its average price appears to follow a pattern that is also positive. As the number of bathrooms increases, the average price tends to go up too, which is a predictable outcome. This trend might help explain why hotel rooms are generally more expensive compared to other types of rentals. Since hotel rooms usually come with a private bathroom, this feature contributes to their higher cost.

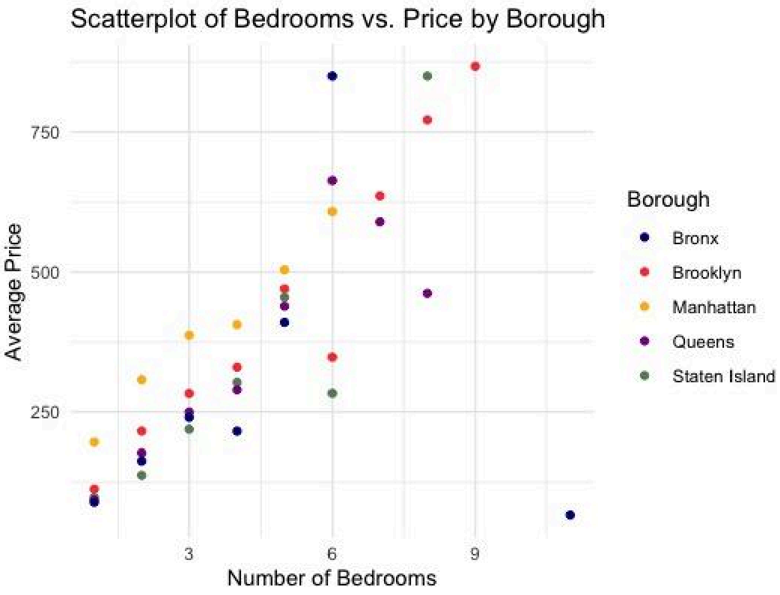
4. **Exploring the relationship between bathroom (shared vs private) and price**

Continuing with the topic of bathrooms, the following graph presents the average prices for accommodations with shared versus private bathrooms. This graph underscores the fact that having a private bathroom significantly increases the cost, often more than doubling it. The presence and type of bathroom prove to be strong indicators of the average price. This graphic was made prior to imputing the dataset, which means that there are NAs present in the dataset. However, we can still conclude that a trend does exist.

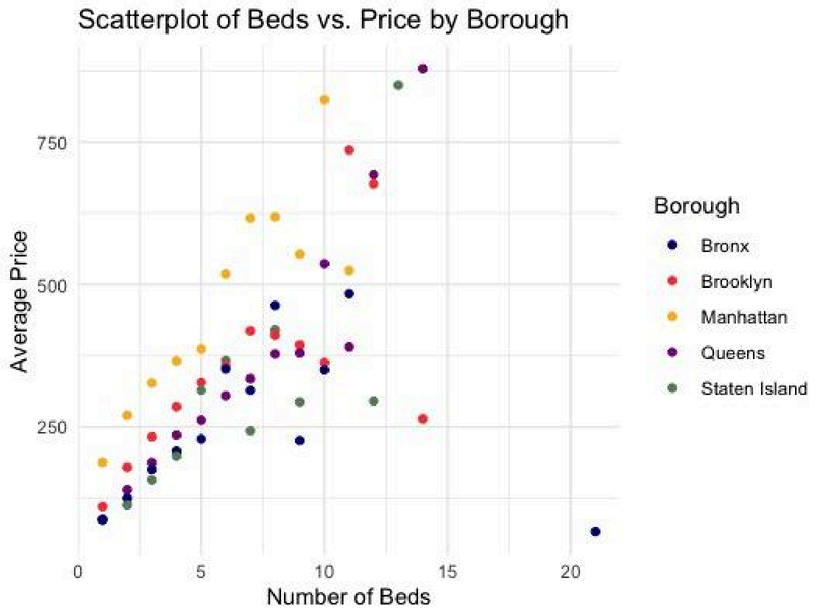


5. **Exploring the relationship between bedrooms and price**

This graph, much like the one comparing accommodation capacity to average price, illustrates the relationship between price and the number of bedrooms in a rental property. It's noteworthy that in Manhattan, there are no data points for rentals with more than six bedrooms. This could be because the cost per square foot in Manhattan is so high that having more than six bedrooms becomes financially impractical.



6. **Exploring the relationship between beds and price**

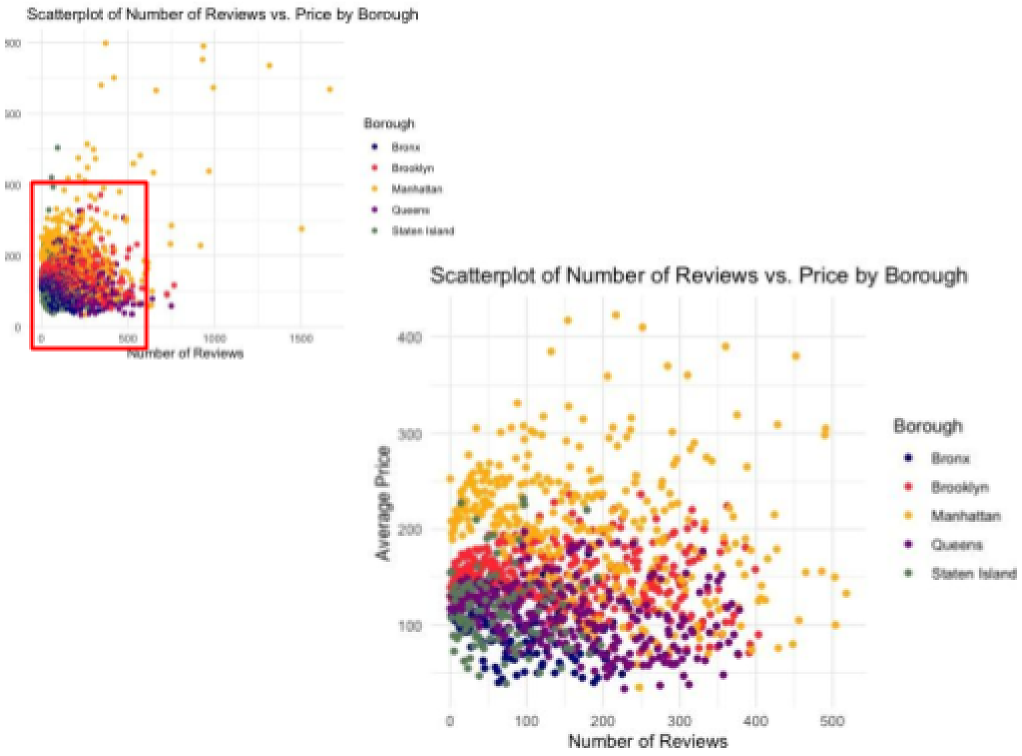


The trend observed is that as the number of beds in a property goes up, so does the price. It would be intriguing to have more detailed information about the types of beds involved. For instance, knowing whether they are queen or king-sized could be important, as these larger beds can accommodate two people instead of one. Additionally, there appears to be a house in the Bronx with 20 beds, which would certainly be interesting to investigate further.

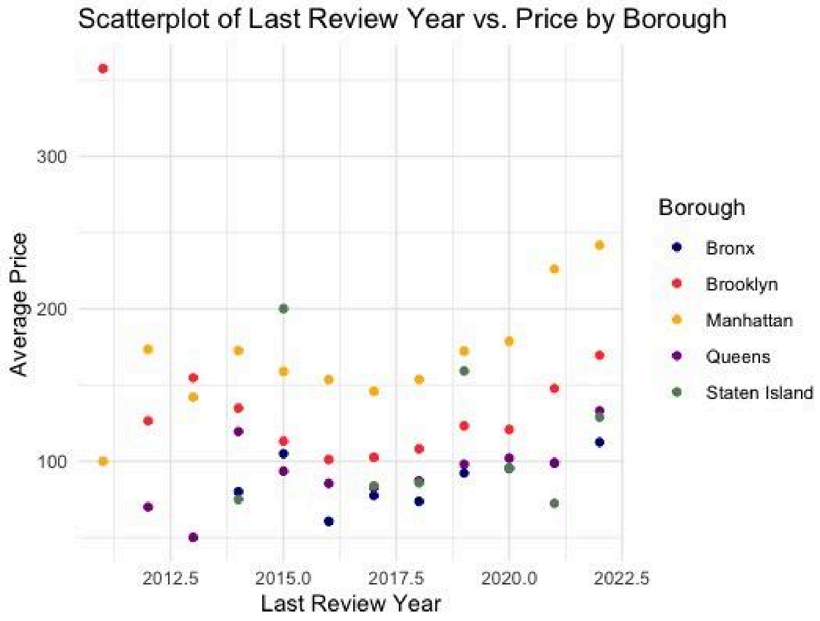
7. **Exploring number of reviews and price**

Analyzing how reviews impact price offers another fascinating perspective. Our initial assumption is that reviews may not significantly influence the price, but rather affect the frequency of bookings.

Upon examining the scatterplot, it becomes evident that there isn't a straightforward correlation between the number of reviews a listing has and its average price.

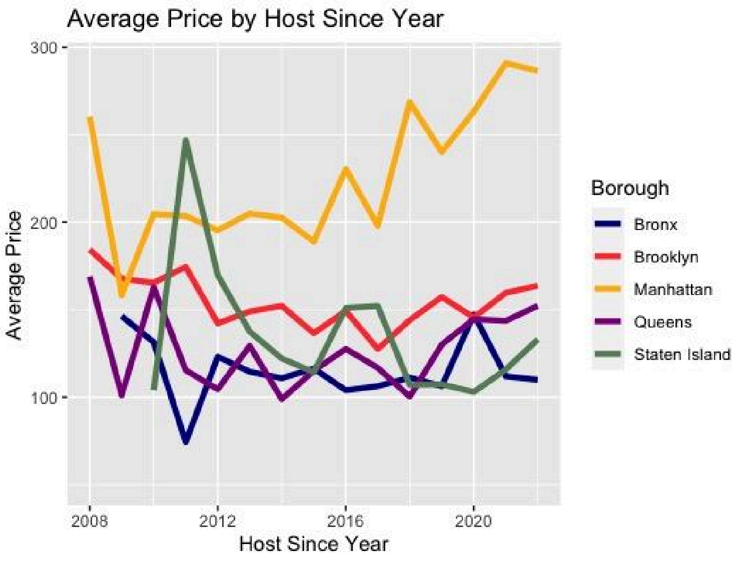


8. **Exploring the relationship between last reviewed and price**



There doesn't appear to be a distinct relationship between the date of the last review and the price. While there is a slight trend suggesting prices might increase with more recent reviews, we cannot definitively conclude that there is a significant impact.

9. **Exploring the relationship between host since and price**



Finally, we explored how the length of time a user has been hosting correlates with the average listing price. The trend is not straightforward to interpret. In Manhattan, for instance, newer hosts appear to be setting higher prices, which could be attributed to rising rent costs or inflation. However, in boroughs like Brooklyn or the Bronx, this relationship is more stable. The duration of a host's experience doesn't seem to significantly affect the average price in these areas.

Answers to Questions task 1

Below are the answers to actual questions asked in Task 1

Which borough has the most rentals? Manhattan → 16607 rentals

1	Manhattan	16607
2	Brooklyn	15567
3	Queens	6495
4	Bronx	1583
5	Staten Island	401

Which room type is the most common? Entire home/apt → 23078

	room_type	count
1	Entire home/apt	23078
2	Hotel room	138
3	Private room	16914
4	Shared room	523

Does a particular borough have a typical rental type?

Manhattan → Entire home/apt

Queens → Private Room

Staten Island → Entire home/apt

Brooklyn → Entire home/apt

Bronx → Private Room

	Neighbourhood_group	room_type	count
1	Bronx	Entire home/apt	715
2	**Bronx**	**Private room**	**826**
3	Bronx	Shared room	42
4	**Brooklyn**	**Entire home/apt**	**8584**
5	Brooklyn	Hotel room	5
6	Brooklyn	Private room	6814
7	Brooklyn	Shared room	164
8	**Manhattan**	**Entire home/apt**	**10632**
9	Manhattan	Hotel room	125
10	Manhattan	Private room	5643
11	Manhattan	Shared room	207
12	Queens	Entire home/apt	2910
13	Queens	Hotel room	8
14	**Queens**	**Private room**	**3470**
15	Queens	Shared room	107
16	**Staten Island**	**Entire home/apt**	**237**

17 Staten Island	Private room	161
18 Staten Island	Shared room	3

Which borough offers the most affordable rentals? Bronx with an average price of 112

1 Bronx	112.
2 Queens	123.
3 Staten Island	129.
4 Brooklyn	149.
5 Manhattan	224.

What geographical features can affect price? Some geographical features that can affect the price are latitude, longitude, neighborhood, and the borough.

Do property types change based on location

This first table provides us with information of what is the most common room type by borough. It also tells us how many listings there are of that room type.

1 Staten Island	Entire home/apt	237
2 Bronx	Private room	826
3 Queens	Private room	3470
4 Brooklyn	Entire home/apt	8584
5 Manhattan	Entire home/apt	10632

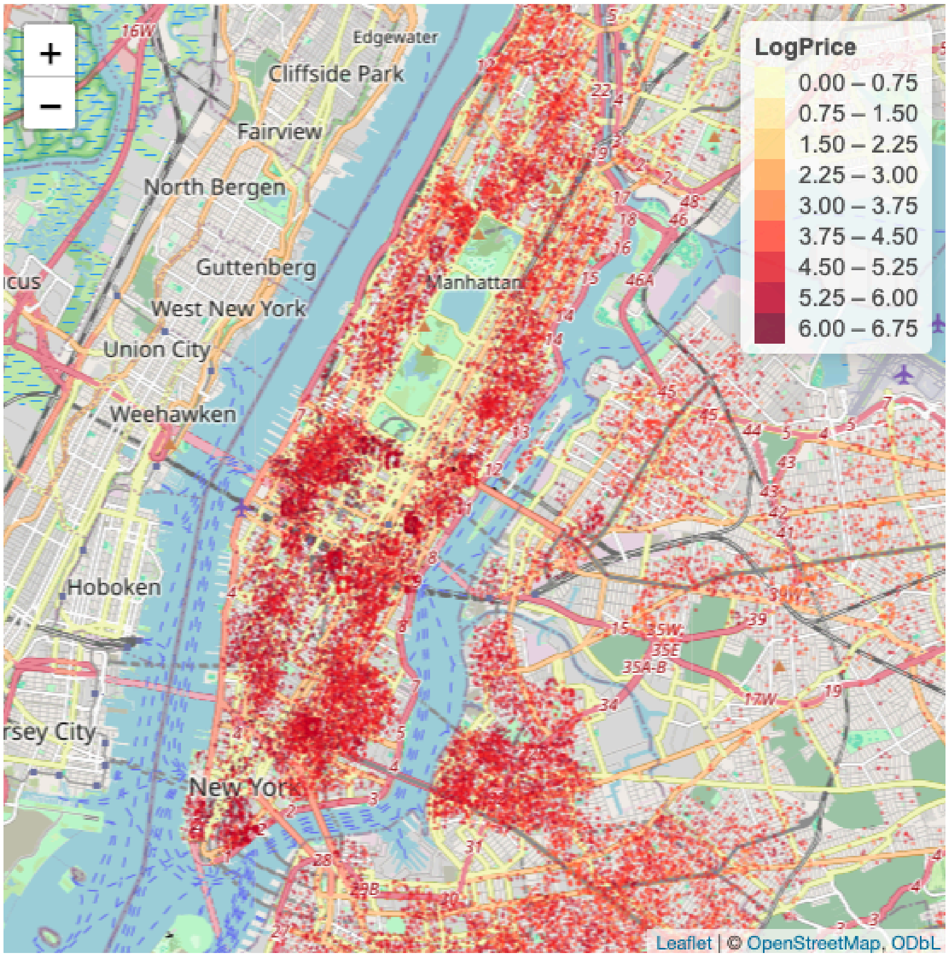
We see it does. Bronx and Queens have more rentals for Private Rooms while Staten Island, Brooklyn and Manhattan have more Entire home/apt rentals.

The second table provides us with information about what is the most common accommodation type by borough (listings for 2 guests are the most common in every borough)

This is by accommodation:

Neighborhood group	Accommodates	Count
1 Staten Island	2	143
2 Bronx	2	670
3 Queens	2	2818
4 Brooklyn	2	6787
5 Manhattan	2	8012

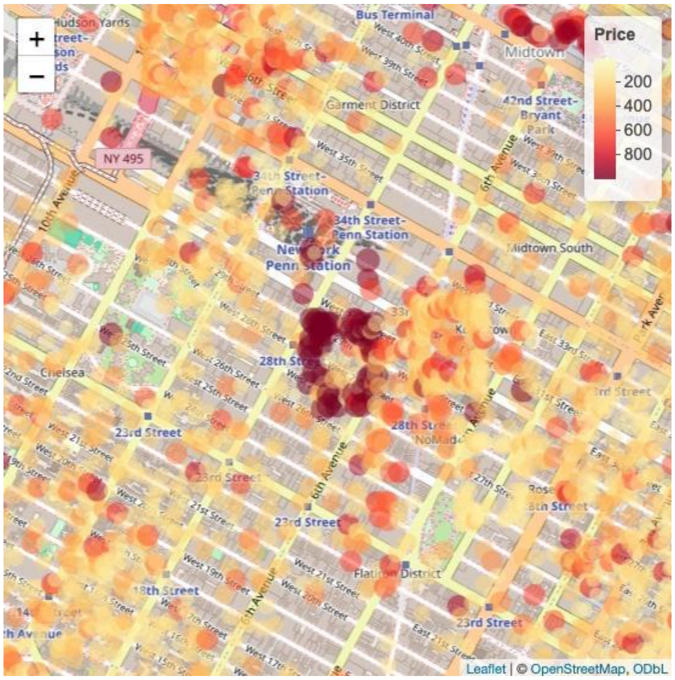
Do property types change based on proximity to points of interest?



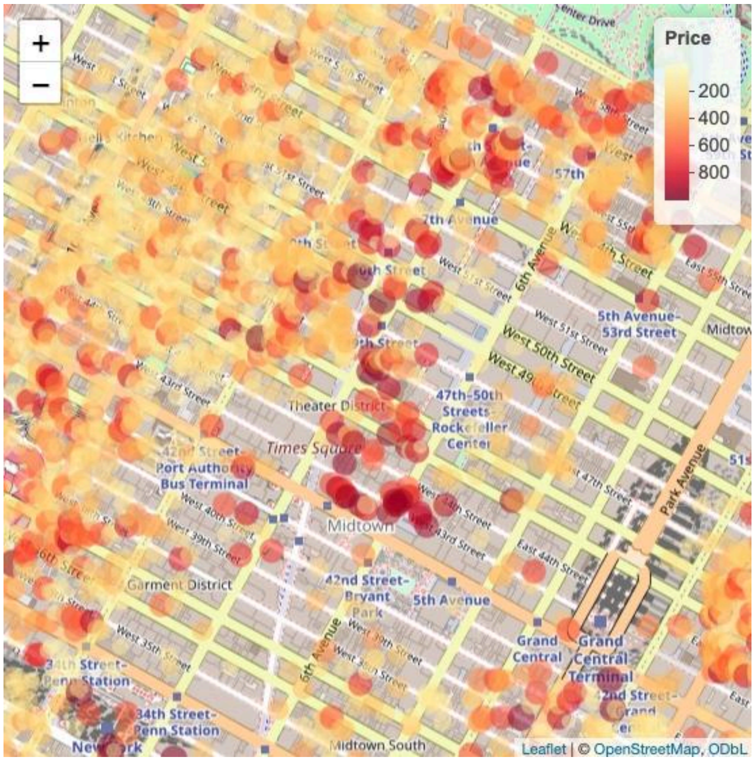
To enhance our understanding of price variations in the heatmap, we applied a logarithmic transformation. This approach makes it easier to discern changes in price, offering a clearer and more defined subgrouping.

This heatmap begins to reveal certain patterns. For those familiar with New York's layout, the Financial District in southern Manhattan stands out with notably high prices. Similarly, the Lower East Side of Central Park also shows elevated prices. As one moves northward through Manhattan, there's a noticeable decrease in prices.

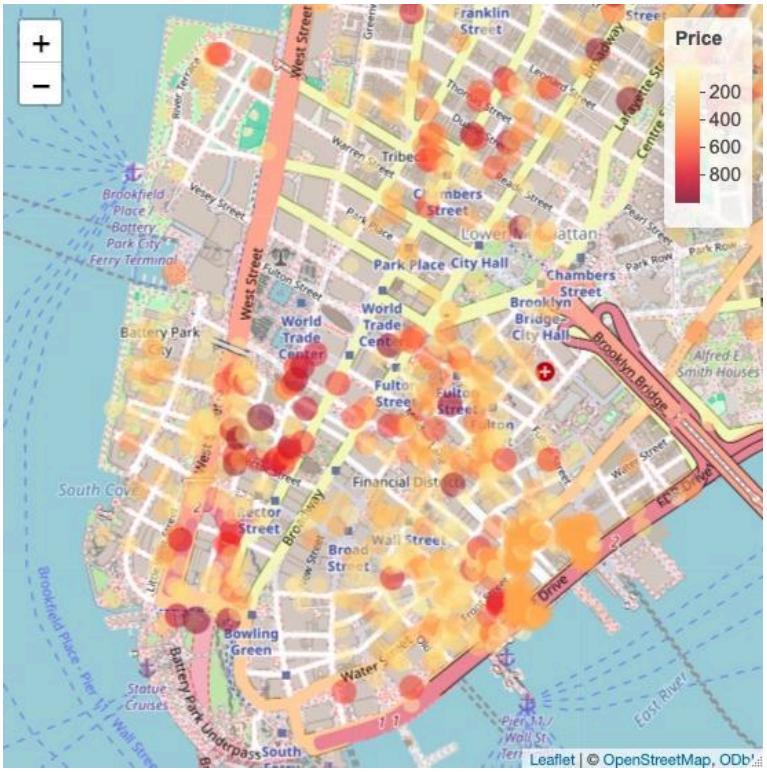
If we subgroup in smaller points of interest we identify the following



The area stretching from Chelsea to Penn Station, a major New York train hub, appears to be on the pricier side. This is likely due to its convenience for frequent travelers.



In the bustling area from the Theater District and Midtown to Times Square and Central Park, including Broadway and Grand Central Terminal, prices are notably high. This zone, rich in tourist attractions, is very popular among visitors. Times Square, one of the most famous squares globally, stands out in the middle of the heatmap with some of the highest prices, reflecting its appeal and prime location

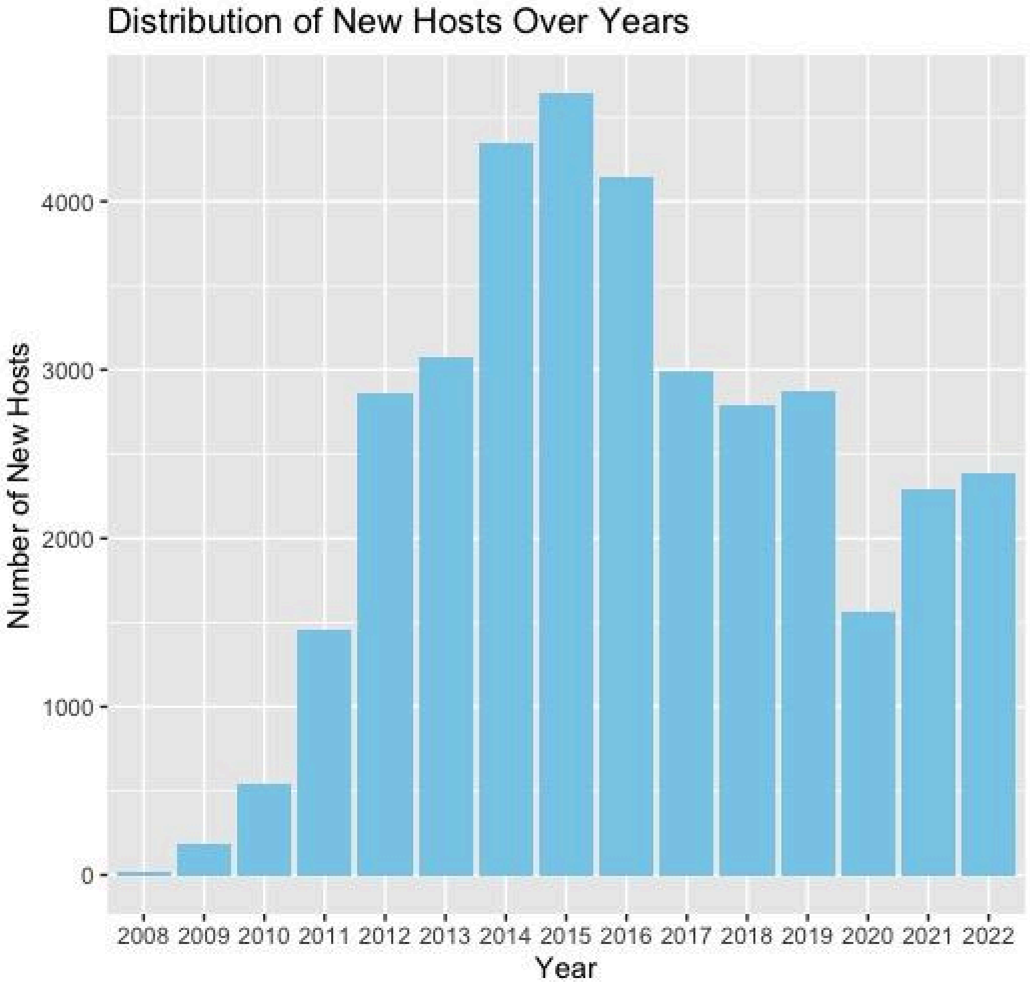


The renowned Financial District in New York, home to Wall Street, the World Trade Center, and the New York Stock Exchange, shows a trend of high prices. The areas around the World Trade Center and Bowling Green, known for their stunning river views, are particularly expensive.

What is the price breakdown for a particular borough?

Neighborhood group	Min price	Max price	Average price	Median price	Total listings
Bronx	10	850	112	89	1583
Queens	10	999	123	92	6495
Staten Island	30	850	129	100	401
Brooklyn	10	999	149	117	15567
Manhattan	10	999	224	170	16607

What year had the largest number of new hosts?



2015 had the largest number of new hosts.

Which host is the most profitable?

- **Some tips: each customer leaves a review after staying, the number of nights for each rental can be assumed to be a single night for revenue calculation purposes**

Most profitable host is Sonder (NYC)
hostID: 158969505

Total Revenue = $3,244,517.00 → assuming the number of nights for each rental is a single night.

How much has TBB earned from NYC rentals?

- **TBB's fee for each transaction is 3% of the listing price**

TBB Earnings = $5,511,939

Task 2

Given the considerable differences among each borough, we decided it would be more precise and yield more analyzable outcomes if we created separate models for each specific borough: Manhattan, Brooklyn, Staten Island, Queens, and The Bronx.
For some boroughs we did subsetting to gather more information and create a better predictor and more accurate model.

Model 1 (MANHATTAN)

a. **No sub-setting**

Linear regression model without subsetting
Predictor variables:

- Accommodates
- Last_review_year
- Room_type
- Bathrooms
- Neighborhood
- Bedrooms
- Beds
- Shared

Adjusted R^2 = 0.4079
RMSE = 134.2256

This specific model doesn't prove to be very useful. Its low adjusted R^2 and less-than-ideal RMSE, with a variation of 134 in price, indicate that it's not particularly effective or beneficial. Therefore, we've decided to improve our approach by subsetting the data.

b. **Sub-setting by neighborhood**

Linear regression model with subsetting by neighborhood

We chose to subset the data by neighborhoods to try and distinguish between the more expensive and cheaper listings. To do this, we focused on the top four neighborhoods in Manhattan with the highest average prices. These are the Theater District, Tribeca, Financial District, and Midtown.

By concentrating on these areas, our goal is to differentiate the higher-priced listings from the more affordable ones.

The following is a list of the top ten neighborhoods and their average prices:

Neighbourhood	**Average Price**
1. Theater District	425
2. Tribeca	359
3. Financial District	326
4. Midtown	309
5. Chelsea	298
6. SoHo	296
7. Flatiron District	277
8. West Village	269
9. NoHo	267
10. Hell’s Kitchen	265

We decided to make a cut at the $300 per night mark for the average price for simplicity and to set a standard claiming that listings > $300 per night are considered expensive and therefore in a “hot” neighborhood.

Hot neighborhoods (listings > $300 per night) model Predictor variables:

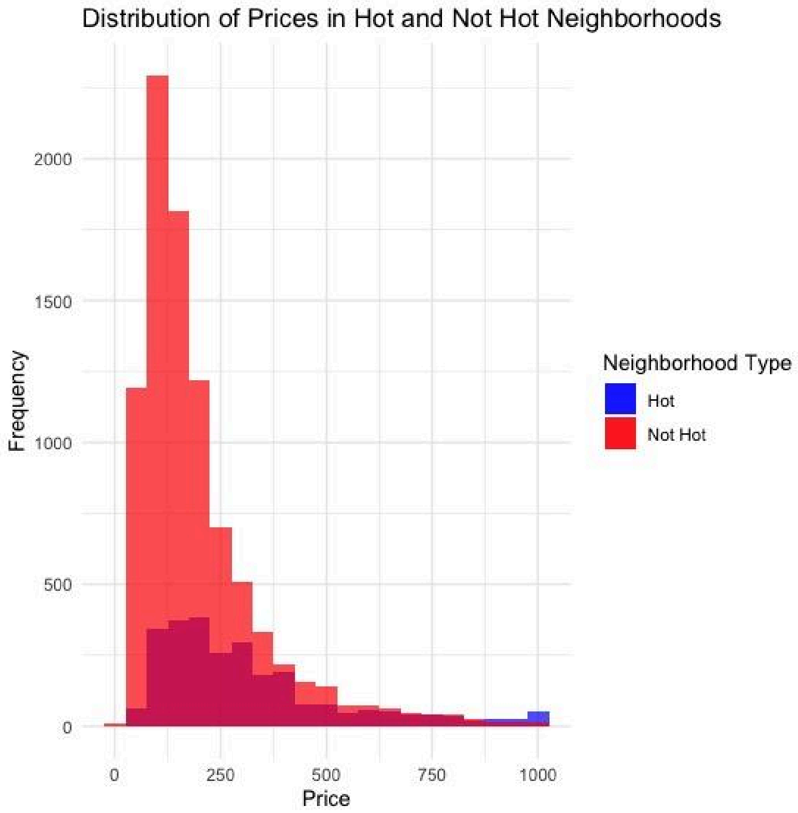
- Accommodates
- Last_review_year
- Shared
- Room_type
- Bathrooms
- Beds
- Neighborhood

Adjusted R^2 = 0.2536
RMSE = 177.0994
This model tells us that predicting the prices of the listings in the hot neighborhoods is difficult. It might be because there exists a lot of variability. In a hot neighborhood such as Midtown (near Time Square) you could find expensive penthouses, but you could also find very small apartments that are very cheap.

Not hot neighborhoods (listings < $300 per night) model Predictor variables:

- Accommodates
- Room_type
- Bathrooms
- Last_review_year
- Shared
- Bedrooms
- Beds
- Neighborhood

Adjusted R^2 = 0.4075
RMSE = 123.3546



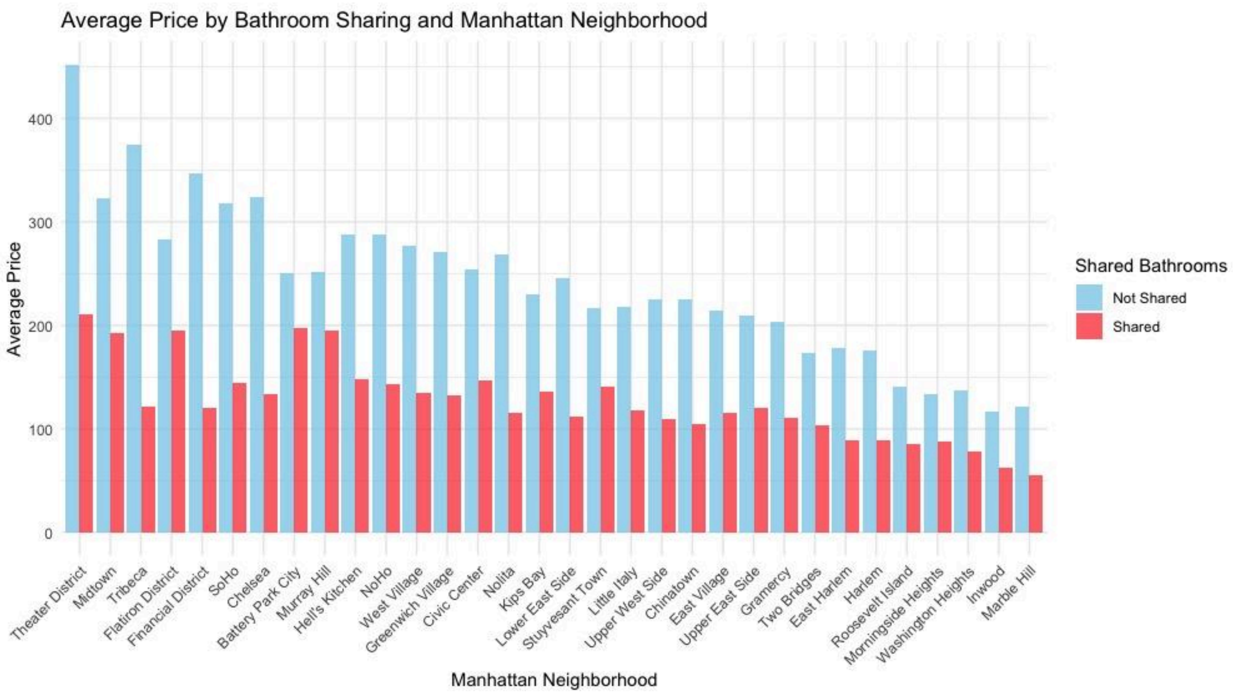
This graph helps us understand what is going on in Manhattan neighborhoods between price and hot vs. not hot neighborhoods. As you can see, hot and not hot neighborhoods in Manhattan follow similar trends, with hot neighborhoods only having relatively more listings with prices above $800 per night.

Since they generally follow a similar trend in terms of price, sub-setting the data by hot and not hot neighborhoods might not be the best approach to take when trying to estimate the prices for listings in Manhattan.

Which is why we make the next model, subsetting the data by shared and not shared bathroom.

c. Subsetting by bathroom Shared Vs. Not Shared

The following model is also for Manhattan, but instead of subsetting the data by hot or not hot neighborhoods, we separated the data by shared and not shared bathrooms.



According to this graph, it is evident that in Manhattan, the average price of a listing greatly differs if a listing has a shared bathroom or does not have a shared bathroom.
With that being said, we believed that taking such an approach would help us separate those more expensive listings from the least expensive listings.

Not shared bathrooms linear regression model predictor variables:

- Accommodates
- Last_review_year
- Room_type
- Bathrooms
- Neighborhood
- Host_id
- Bedrooms
- Beds

Adjusted R^2 = 0.3626
RMSE = 143.4244

Shared bathrooms linear regression model predictor variables:

- Accommodates
- Last_review_year
- Room_type
- Bathrooms
- Beds

- Neighborhood

Adjusted $R^2 = 0.2232$
RMSE = 71.3344

CONCLUSION ALL MANHATTAN MODELS
Overall between the different models for predicting prices in Manhattan, when subsetting the data (the cheaper/less desirable listings) tend to be predicted with a lower RMSE.

We believe that subsetting the data for a borough such as Manhattan is necessary because there exists so much variability in the types of listings that are present. Even though the RMSE for the model without subsetting (1a) has an RMSE of ~$134, it would not be the best idea to apply this model to the evaluation set.

We believe that subsetting the data based on whether bathrooms are shared or not (1c) turned out to be more effective than dividing it by neighborhood (1b), as indicated by the lower RMSE in the models.

Moreover, this approach of segregating based on shared and not shared bathrooms allowed us to develop two distinct models for Manhattan, instead of just one. This is particularly advantageous because Manhattan has a high concentration of listings compared to other areas. By specifically targeting shared and non-shared bathroom listings in our training set, we can more accurately predict prices for these two categories.

For instance, when it comes to testing, it's much more beneficial to have separate models for shared and non-shared bathrooms in Manhattan. This approach can yield a decent prediction for non-shared and an even stronger prediction for shared bathrooms. In contrast, a single model covering all Manhattan listings might only provide moderately satisfactory predictions across the board. This finer segmentation simply leads to better, more tailored predictive outcomes.

Model 2 (BROOKLYN)
The next models correspond to making predictions for Brooklyn. Brooklyn also has a lot of listings (15,567), so we decided to take a similar approach by subsetting the data based on bathrooms. But we must see all the models.

a. **No subsetting**

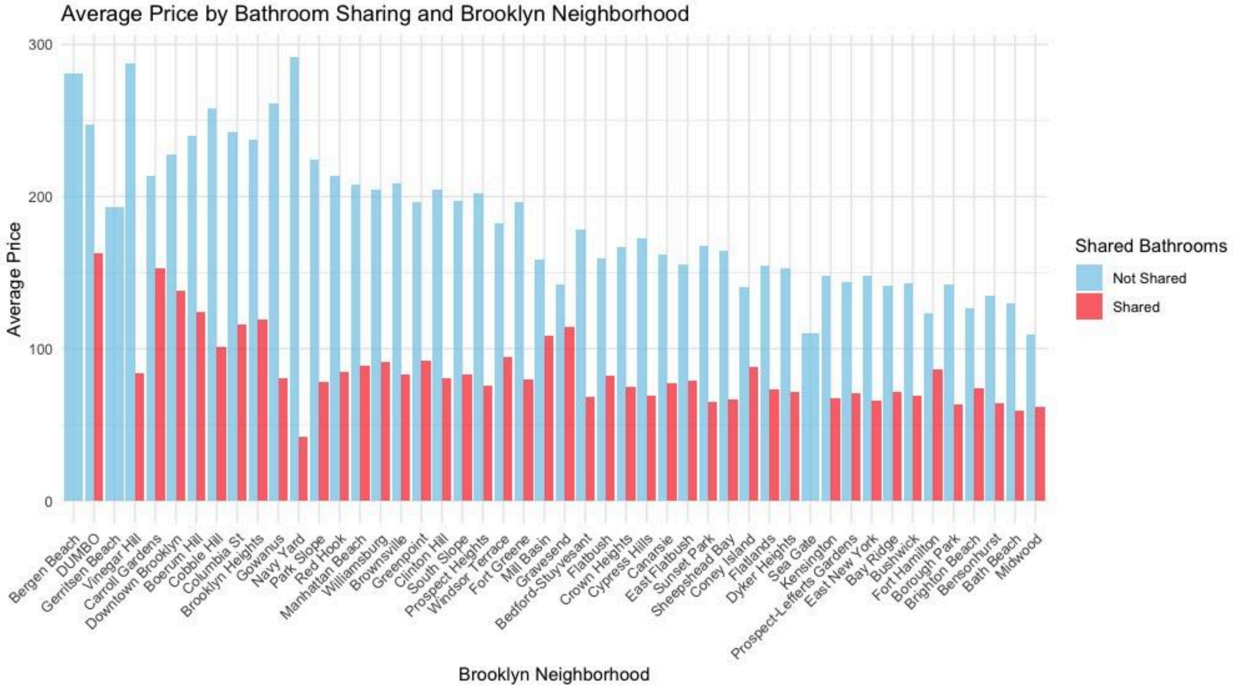
Linear regression model without subsetting
Predictor variables:

- Accommodates

- Room_type
- Bathrooms
- Shared
- Bedrooms
- Beds
- Last_review_year
- Number_of_reviewss
- Longitude
- Latitude
- Neighbourhood

Adjusted R^2 = 0.5223
RMSE = 77.3756
This model has shown the most promising results so far. Despite the large number of listings, it achieves a relatively low RMSE, making it a pretty good model.

b. Subsetting by bathroom Shared Vs. Not Shared



According to this graphic, similarly to Manhattan, Brooklyn also follows a similar trend in the sense that shared bathrooms have a relatively lower average price. It is important to take this into account when making our predictive models. With that being said, we will take a similar approach. We also decided to apply a similar approach because Brooklyn has a lot of listings (15567) to be exact.

Linear regression model by subsetting based on not shared bathrooms

Predictor Variables for not shared bathrooms :

- Accommodates
- Room_type
- Bathrooms
- Bedrooms
- Beds
- Last_review_year
- Number_of_reviews
- Longitude
- Latitude
- Neighborhood

Adjusted R^2 = 0.4934
RMSE = 85.65703

Linear regression model by subsetting based on shared bathrooms

Predictor Variables:

- Accommodates
- Room_type
- Bathrooms
- Last_review_year
- Longitude
- Latitude

Adjusted R^2 = 0.1158
RMSE = 43.25507

FINAL CONCLUSION BROOKLYN
Similar to the manhattan data, we took an approach that separates the listings by shared and not shared bathrooms. This is because there are many listings in Brooklyn and a lot of variability in the prices. Furthermore, having 2 models to predict Brooklyn will perform better than one model on the evaluation set.

Model 3 (STATEN ISLAND)
For the rest of the models we decided to just use the whole borough and not subset them because they have significantly less listings. Subsetting the data is not necessary and would not have enough for splitting the good training & testing sets.

a. **Linear regression model staten island (no subsetting)**

Predictor variables:

- Accommodates
- Bathrooms
- Beds
- Number_of_reviews

Adjusted R^2 = 0.5413
RMSE = 55.42538

Model 4 (QUEENS)

a. **Linear regression model queens (no subsetting)**

Predictive variables:

- Accommodates

- Room_type
- Bathrooms
- Bedrooms
- Last_review_year
- Number_of_reviews
- Neighborhood

Adjusted R^2 = 0.5235
RMSE = 65.08862

Model 5 (BRONX)

a. **Linear regression model queens (no subsetting)**

Predictor variables:

- Accommodates
- Room_type
- Bathrooms
- Bedrooms
- Last_review_year
- Number_of_reviews
- Longitude
- Latitude
- Neighborhood

Adjusted R^2 = 0.4456
RMSE = 54.26868

Final Recommendation

Based on the extensive analysis and comparisons of the various models for each borough in New York City, the final recommendation for the best predictive model depends on the specific characteristics and needs of each borough.

For Manhattan, the most effective approach appears to be subsetting the data by whether bathrooms are shared or not. This method yielded two distinct models – one for listings with shared bathrooms and another for those with not shared bathrooms. The key advantage of this approach is its ability to mold to the diverse range of listings in Manhattan, characterized by a high concentration and variability. Specifically, the model focusing on shared bathrooms showed a more accurate prediction with a lower RMSE, indicating its effectiveness in capturing the price dynamics in this segment. In contrast, a single model encompassing all listings in Manhattan would likely provide only moderately satisfactory predictions.

In Brooklyn, the model without subsetting, which includes a broad range of predictor variables like Accommodates, Room_type, Bathrooms, and geographical coordinates, proved to be the most promising, achieving a relatively low RMSE despite the large number of listings. This suggests that a comprehensive model is sufficient for capturing the price trends in Brooklyn, possibly due to less variability in listing types compared to Manhattan.

For Staten Island, Queens, and the Bronx, where the number of listings is significantly lower, models without subsetting were found to be sufficient. These models successfully capture the price dynamics without the need for further segmentation, as evidenced by their respectable adjusted R-squared values and RMSEs. This simplicity is likely due to less diversity and lower variability in listings within these boroughs.

In conclusion, while a one-size-fits-all model might seem appealing, the unique characteristics of each borough necessitate tailored approaches. For Manhattan and Brooklyn, where the listing numbers and variability are high, specialized models (either by subsetting or including a wide range of variables) are more effective. In contrast, for Staten Island, Queens, and the Bronx, a more generalized model approach works. This tailored strategy ensures more accurate and reliable predictions across the diverse landscape of New York City's ToreroBnB (TBB) market.