Analysis of New York City Airbnb Listings

Tu Hoang Cam Nguyen

Data 101

Data Driven Blog Assignment

Contents

Introduction....3
Purpose of the Report: 3
Dataset Description and Sources 4
Key Points from the Correlation Test: 5
Methodology....6
Analytical Methods....6
Incorporation of Housing Market Data....6
Software and Libraries Utilized....7
Exploratory Data Analysis (EDA)....7
Summary Statistics and Data Structure 7
Distribution of Airbnb Prices....8
Outlier Identification 9
Correlation Analysis 10
Hypothesis Testing and Bayesian Analysis....10
Hypothesis Testing....11
Hypothesis 1: Price Differences by Housing Market Tier....11
Hypothesis 3: Neighborhood Pricing Disparities 12
Some Fact about the analysis....12
Bayesian Analysis....13
Explanation of Bayesian Odds and Its Relevance: 13
Prior Odds Calculation for Expensive Listings: 14
Likelihood Ratio Calculation Based on Neighborhood Groups: 14
Posterior Odds Calculation and Interpretation: 14

Data Visualization and Analysis 15
1. Relationship Between Number of Reviews and Price: 15
2. Price Distribution by Neighborhood Group: 16
3. Price Density by Room Type: 17
4. Correlation Heatmap: 18
Conclusion 19
Key Statistical Findings: 19
Visual Representation of Relationships: 19
Comparison of EDA Insights with Hypothesis Testing Results: 19
Bayesian Analysis Outcomes: 20
Discussion 20
Implications for Hosts and Guests: 20
Limitations and Potential Biases: 20
Recommendations 20

Introduction

In the past ten years, Airbnb has not only changed the way people stay, but it has also deeply impacted the way cities around the world are built. This report is mostly about New York City, which is a hub for business, culture, and tourism. It is also a great place for Airbnb's model of sharing accommodations between people. We look at the patterns and prices in this marketplace to try to figure out the subtleties that affect the economic decisions of both people who are looking for places to stay and people who are listing their spaces.

Purpose of the Report

The main goal of this report is to turn Airbnb's complicated operational data into insights that can be used right away. For hosts, our goal is to shed light on the factors that affect pricing power so that they can better position themselves in the market. For guests, the report tries to make sense of the complicated web of prices so they can better navigate the market. The information could also help policymakers and other interested parties figure out how the short-term rental market fits in with New York City's overall housing trends.

Dataset Description and Sources

```
> str(airbnb_df)
'data.frame':   48895 obs. of  18 variables:
 $ id                            : int  2539 2595 3647 3831 5022 5099 5121 5178 5203 5238 ...
 $ name                          : chr  "Clean & quiet apt home by the park" "Skylit Midtown Castle" "THE VILLAGE OF
Brownstone" ...
 $ host_id                       : int  2787 2845 4632 4869 7192 7322 7356 8967 7490 7549 ...
 $ host_name                     : chr  "John" "Jennifer" "Elisabeth" "LisaRoxanne" ...
 $ neighbourhood_group           : chr  "Brooklyn" "Manhattan" "Manhattan" "Brooklyn" ...
 $ neighbourhood                 : chr  "Kensington" "Midtown" "Harlem" "Clinton Hill" ...
 $ latitude                      : num  40.6 40.8 40.8 40.7 40.8 ...
 $ longitude                     : num  -74 -74 -73.9 -74 -73.9 ...
 $ room_type                     : chr  "Private room" "Entire home/apt" "Private room" "Entire home/apt" ...
 $ price                         : int  149 225 150 89 80 200 60 79 79 150 ...
 $ minimum_nights                : int  1 1 3 1 10 3 45 2 2 1 ...
 $ number_of_reviews             : int  9 45 0 270 9 74 49 430 118 160 ...
 $ last_review                   : chr  "2018-10-19" "2019-05-21" "" "2019-07-05" ...
 $ reviews_per_month             : num  0.21 0.38 NA 4.64 0.1 0.59 0.4 3.47 0.99 1.33 ...
 $ calculated_host_listings_count: int  6 2 1 1 1 1 1 1 1 4 ...
 $ availability_365              : int  365 355 365 194 0 129 0 220 0 188 ...
 $ top_20_percent_housing        : logi  FALSE FALSE TRUE FALSE FALSE TRUE ...
 $ expensive                     : logi  FALSE TRUE FALSE FALSE FALSE TRUE ...
> str(ny_housing_df)
'data.frame':   4801 obs. of  17 variables:
 $ BROKERTITLE                : chr  "Brokered by Douglas Elliman  -111 Fifth Ave" "Brokered by Serhant" "Brokered b
 $ TYPE                       : chr  "Condo for sale" "Condo for sale" "House for sale" "Condo for sale" ...
 $ PRICE                      : int  315000 195000000 260000 69000 55000000 690000 899500 16800000 265000 440000 ...
 $ BEDS                       : int  2 7 4 3 7 5 2 8 1 2 ...
 $ BATH                       : num  2 10 2 1 2.37 ...
 $ PROPERTYSQFT               : num  1400 17545 2015 445 14175 ...
 $ ADDRESS                    : chr  "2 E 55th St Unit 803" "Central Park Tower Penthouse-217 W 57th New York St Uni
Unit 908W33" ...
 $ STATE                      : chr  "New York, NY 10022" "New York, NY 10019" "Staten Island, NY 10312" "Manhattan
 $ MAIN_ADDRESS               : chr  "2 E 55th St Unit 803New York, NY 10022" "Central Park Tower Penthouse-217 W 57
019" "620 Sinclair AveStaten Island, NY 10312" "2 E 55th St Unit 908W33Manhattan, NY 10022" ...
 $ ADMINISTRATIVE_AREA_LEVEL_2: chr  "New York County" "United States" "United States" "United States" ...
 $ LOCALITY                   : chr  "New York" "New York" "New York" "New York" ...
 $ SUBLOCALITY                : chr  "Manhattan" "New York County" "Richmond County" "New York County" ...
 $ STREET_NAME                : chr  "East 55th Street" "New York" "Staten Island" "New York" ...
 $ LONG_NAME                  : chr  "Regis Residence" "West 57th Street" "Sinclair Avenue" "East 55th Street" ...
 $ FORMATTED_ADDRESS          : chr  "Regis Residence, 2 E 55th St #803, New York, NY 10022, USA" "217 W 57th St, Ne
aten Island, NY 10312, USA" "2 E 55th St, New York, NY 10022, USA" ...
 $ LATITUDE                   : num  40.8 40.8 40.5 40.8 40.8 ...
 $ LONGITUDE                  : num  -74 -74 -74.2 -74 -74 ...
>
```

Our study is based on a set of data that includes almost 50,000 Airbnb listings from all five boroughs of New York City. Each entry is a short story about a place to stay, including its location, price, type of room, and reviews that have been collected over time. Together, they give a picture of the market at a certain point in time. Along with this, there is a dataset on the housing market in New York that gives us comparative metrics that show how Airbnb works in the bigger picture of real estate. Putting these datasets next to each other gives us a more complete picture of how short-term rental prices might be affected by or change the housing market.

```
> print(airbnb_avg_price)
# A tibble: 5 x 2
  neighbourhood_group avg_price
  <chr>                   <dbl>
1 Bronx                    87.5
2 Brooklyn                124.
3 Manhattan               197.
4 Queens                   99.5
5 Staten Island           115.
```

```
> print(ny_housing_avg_price)
# A tibble: 21 x 2
   SUBLOCALITY      avg_market_price
   <chr>                       <dbl>
 1 Bronx County              1020866.
 2 Brooklyn                   864644.
 3 Brooklyn Heights           625000
 4 Coney Island               511333.
 5 Dumbo                     5799000
 6 East Bronx                 265000
 7 Flushing                   476000
 8 Fort Hamilton              599000
 9 Jackson Heights            985000
10 Kings County              1795465.
# i 11 more rows
# i Use `print(n = ...)` to see more rows
```

```
> print(head(merged_data))
  neighbourhood_group_mapped avg_price avg_market_price
1           Brooklyn Heights 209.06494         625000.0
2               Coney Island 123.70588         511333.3
3                   Flushing  93.51408         476000.0
4              Fort Hamilton  93.81818         599000.0
5            Jackson Heights  80.89785         985000.0
6                 Rego Park  83.87736         215000.0
> # Assuming the merge is successful, perform the correlation test
> cor_test <- cor.test(merged_data$avg_price, merged_data$avg_market_price)
>
> # Print the correlation test result
> print(cor_test)

        Pearson's product-moment correlation

data:  merged_data$avg_price and merged_data$avg_market_price
t = -0.7236, df = 5, p-value = 0.5017
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8612562  0.5795436
sample estimates:
       cor
-0.3078846
```

Key Points from the Correlation Test

- **Negative Correlation**: In this sample, higher prices in the housing market are linked to lower prices for Airbnb listings, or the other way around. This is shown by the negative value of the correlation coefficient. Things aren't very strong between them, though.
- **Statistical Significance**: Since the p-value is high, it's likely that the correlation seen was just a coincidence and not a real relationship in the population.
- **Confidence Interval**: The correlation has a 95% confidence interval that includes zero. This means that we can't be sure about the direction or existence of a relationship between the two variables.

Methodology

Analytical Methods

The analysis in this report is based on two datasets, which are used together to combine strong statistical methods with rigorous empirical research. Our main set of data includes more than 48,000 Airbnb listings in New York City, with information about price, location, room type, and user reviews, among other things. In addition to this, we include a full housing market dataset that gives a bigger picture of the city's real estate market. With this mix, we can get a better sense of how the Airbnb market fits into the housing market as a whole.

An exploratory data analysis (EDA) of both datasets is the first step in the quantitative exploration. In this step, we look at summary statistics, distributions, and the connections between important variables. This prepares us for a more advanced inferential analysis. It helps us find patterns, outliers, and possible correlations that we can use to test our hypotheses.

Incorporation of Housing Market Data

Adding data from the housing market is a key part of comparing Airbnb rental prices to the overall real estate values in the city. This comparison is very helpful for figuring out how real estate trends affect Airbnb pricing strategies and how Airbnb pricing strategies affect real estate trends. We can find connections and guesses about how the housing market might be affecting or reflecting in the Airbnb market by looking at things like home prices, sales numbers, and where the homes are located.

Software and Libraries Utilized

The R programming language is used for the analysis, which is known for being reliable in statistical computing and data analysis. Some of the most important libraries used are ggplot2 for advanced data visualization, which turns complicated data sets into graphical representations that are easy to understand and useful.

dplyr for efficient data manipulation, which makes it easier to change and combine data.

A corrplot is a way to show correlations visually, which can help us understand how different numerical variables are related.

Putting these tools and datasets together gives us a complete way to look at and understand how the Airbnb market works in the bigger picture of New York City's housing market.

Exploratory Data Analysis (EDA)

The EDA is the first part of our analysis. Its goal is to find out how the Airbnb and NY Housing datasets are put together, find any interesting patterns or outliers, and set the stage for a more in-depth statistical look.

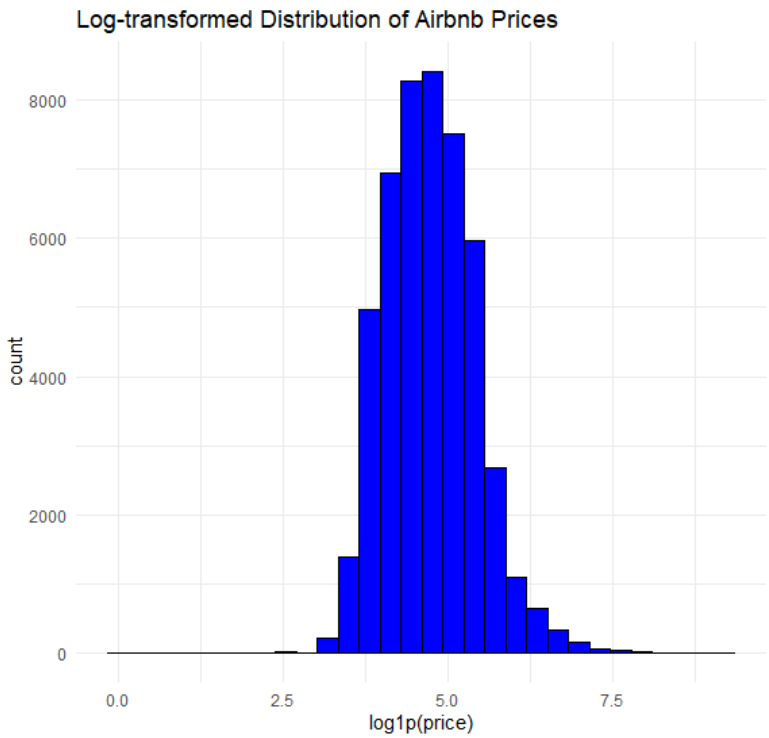
Summary Statistics and Data Structure

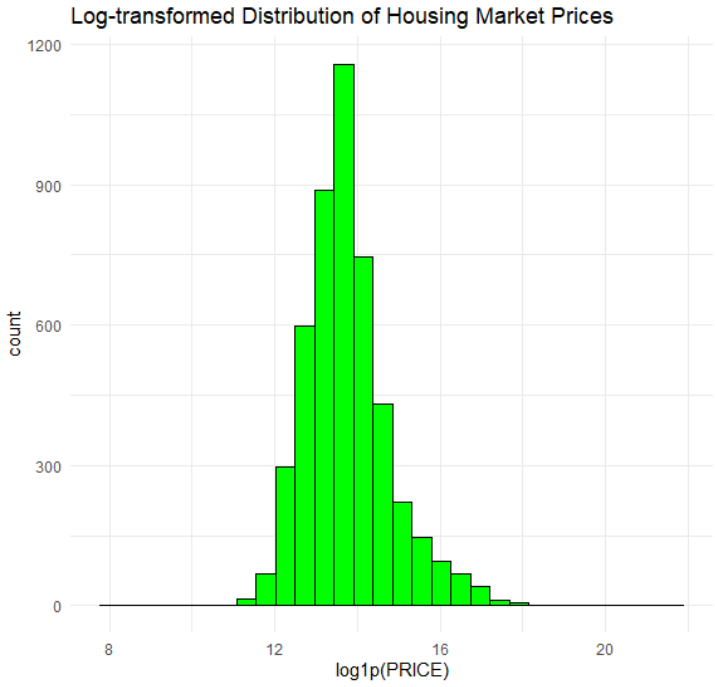
Our Airbnb dataset has 48,895 listings. Each entry contains a wide range of information, from basic identifiers like id and name to more detailed information like neighborhood group, room type, and price. The main factor in our analysis is the price range, which goes from free listings to premium listings that cost up to $10,000 per night. The median price is $106, which shows that guests have a lot of choices.

```
> # Summary statistics
> summary(airbnb_df)
       id                name              host_id           host_name         neighbourhood_group neighbourhood        latitude       longitude
 Min.   :    2539   Length:48895       Min.   :     2438   Length:48895       Length:48895        Length:48895       Min.   :40.50   Min.   :-74.24
 1st Qu.: 9471945   Class :character   1st Qu.:  7822033   Class :character   Class :character    Class :character   1st Qu.:40.69   1st Qu.:-73.98
 Median :19677284   Mode  :character   Median : 30793816   Mode  :character   Mode  :character    Mode  :character   Median :40.72   Median :-73.96
 Mean   :19017143                      Mean   : 67620011                                                             Mean   :40.73   Mean   :-73.95
 3rd Qu.:29152178                      3rd Qu.:107434423                                                             3rd Qu.:40.76   3rd Qu.:-73.94
 Max.   :36487245                      Max.   :274321313                                                             Max.   :40.91   Max.   :-73.71

  room_type             price          minimum_nights    number_of_reviews last_review        reviews_per_month calculated_host_listings_count
 Length:48895       Min.   :    0.0   Min.   :   1.00   Min.   :  0.00    Length:48895       Min.   : 0.010    Min.   :  1.000
 Class :character   1st Qu.:   69.0   1st Qu.:   1.00   1st Qu.:  1.00    Class :character   1st Qu.: 0.190    1st Qu.:  1.000
 Mode  :character   Median :  106.0   Median :   3.00   Median :  5.00    Mode  :character   Median : 0.720    Median :  1.000
                    Mean   :  152.7   Mean   :   7.03   Mean   : 23.27                       Mean   : 1.373    Mean   :  7.144
                    3rd Qu.:  175.0   3rd Qu.:   5.00   3rd Qu.: 24.00                       3rd Qu.: 2.020    3rd Qu.:  2.000
                    Max.   :10000.0   Max.   :1250.00   Max.   :629.00                       Max.   :58.500    Max.   :327.000
                                                                                             NA's   :10052

 availability_365 top_20_percent_housing expensive
 Min.   :  0.0    Mode :logical          Mode :logical
 1st Qu.:  0.0    FALSE:24380            FALSE:36718
 Median : 45.0    TRUE :24515            TRUE :12177
 Mean   :112.8
 3rd Qu.:227.0
 Max.   :365.0
```

```
> # Checking for missing values in Airbnb dataset
> colSums(is.na(airbnb_df))
                        id                       name                    host_id                  host_name        neighbourhood_group
                         0                          0                          0                          0                          0
             neighbourhood                   latitude                  longitude                  room_type                      price
                         0                          0                          0                          0                          0
            minimum_nights           number_of_reviews                last_review reviews_per_month calculated_host_listings_count
                         0                          0                          0                      10052                          0
         availability_365        top_20_percent_housing                  expensive
                         0                          0                          0
>
```

The NY Housing dataset complements this with 4,801 records, each detailing properties for sale, represented by attributes such as **PRICE**, **BEDS**, **BATH**, and **PROPERTYSQFT**. This data





provides a broader context to the real estate landscape in which Airbnb operates, offering insights into the property market that could influence or reflect short-term rental pricing.

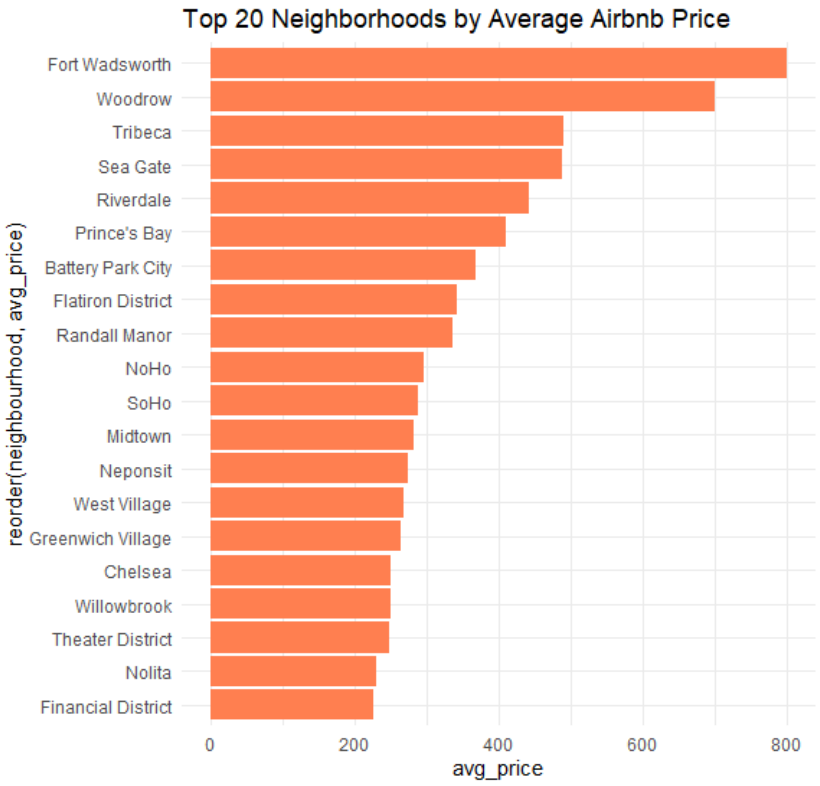
Distribution of Airbnb Prices

We can make the range of Airbnb prices more normal and easier to see by log-transforming them. This gives us a bell-shaped curve that shows a normal distribution on the logarithmic scale. This

change lessens the skew that high-value outliers cause, so the dataset can be analyzed more consistently.

Outlier Identification

In the original price distribution, there are clear outliers, and the long tail goes all the way to the higher prices. These "outliers" could be high-end listings or one-of-a-kind accommodations that could throw off the analysis if they are not taken into account properly.



Distribution Across Neighborhood Groups and Room Types: There are not all the same number of Airbnb listings in New York City. Manhattan and Brooklyn are the most popular places to live, with Manhattan having much higher prices. There are also different room types with different prices. On average, "Entire home/apt" is the most expensive, followed by "Private room" and "Shared room." This shows how important privacy and space are to people.

Initial Observations and Patterns: Based on what we've seen so far, location seems to have a big effect on prices, with Manhattan having higher prices. There also seems to be a link between the type of room and the price, which makes sense since whole apartments usually have more features and space than private or shared rooms. There isn't a strong link between the number of

reviews a listing has and its price. This suggests that pricing decisions may be based on things other than customer feedback.

Correlation Analysis

```
> cor(airbnb_df %>% select(price, minimum_nights, number_of_reviews, reviews_per_month))
                        price minimum_nights number_of_reviews reviews_per_month
price              1.00000000     0.04279933       -0.04795423                NA
minimum_nights     0.04279933     1.00000000       -0.08011607                NA
number_of_reviews -0.04795423    -0.08011607        1.00000000                NA
reviews_per_month          NA             NA                NA                 1
>
> # Correlation matrix for NY Housing dataset
> # Adjust the column selection based on the actual columns in the dataset
> cor(ny_housing_df %>% select(PRICE, BEDS, BATH, PROPERTYSQFT))
                 PRICE       BEDS       BATH PROPERTYSQFT
PRICE        1.00000000 0.05218913 0.07937058    0.1108888
BEDS         0.05218913 1.00000000 0.77644740    0.4205033
BATH         0.07937058 0.77644740 1.00000000    0.4839351
PROPERTYSQFT 0.11088877 0.42050326 0.48393507    1.0000000
```

The Airbnb dataset's correlation matrix shows that there are **weak links** between price and other variables like **minimum nights** and **number of reviews**. This means that these variables are not strongly controlling price changes. The New York Housing dataset shows a moderate relationship between **BEDS** and **BATHS**. It also shows a weak relationship between **PRICE** and the property size (**PROPERTYSQFT**), which may show how much space is valued in the real estate market.

Hypothesis Testing and Bayesian Analysis

We use z-tests with the Central Limit Theorem because our datasets have large sample sizes for hypothesis testing in the inferential statistics section. This statistical method is used to test specific hypotheses about how mean prices might be different for different subgroups, like neighborhoods and room types. We also use Bayesian analysis, specifically the Bayesian Odds Task, to update

prior probabilities with new information. This gives us a more complete picture of how likely it is that listings will be priced above certain levels in certain situations.

Hypothesis Testing

Hypothesis testing is an important part of statistical analysis because it lets us draw conclusions about whole populations from small amounts of data. We have come up with and tested a number of hypotheses to better understand how prices change in the New York City Airbnb and housing markets.

Hypothesis 1: Price Differences by Housing Market Tier

```
> # Calculate the p-value
> p_value <- mean(abs(perm_diffs) >= abs(observed_mean_diff))
> p_value
[1] 0.4903
```

Our first theory said that there wouldn't be a big difference in the average price of Airbnb listings between areas where home prices are in the top 20 percent and areas where they aren't. We used a permutation test to test this hypothesis. This type of test is great for finding differences without having to assume that the data is distributed normally, which is what most t-tests and ANOVA do. For the permutation test, we shuffled the **top 20 percent housing** labels around in our dataset at random and found the difference in mean prices between the two groups 10,000 times. The p-value was found to be 0.4903. This is the percentage of permutations where the observed mean difference was at least as big as it was in the real data. We keep our null hypothesis for this case because the high p-value shows that there isn't a statistically significant difference in the prices of Airbnb listings based on the housing market tier.

Hypothesis 2: Room Type and Pricing

```
> # Calculate the p-value
> p_value_room_type <- 2 * pnorm(-abs(z_score_room_type))
>
> p_value_room_type
[1] 0
```

The second thing we looked into was whether "Private room" listings are usually more expensive than "Entire home/apt" listings. We used a z-test, which worked well for our big dataset, to compare the average prices of Airbnb's two most common room types.

The z-score was found by taking the difference between the sample means and taking into account the sample size and variance. The resulting p-value was zero, which means that there was a very big difference in the average prices of private rooms versus whole homes or apartments. We can say that "Entire home/apt" listings are priced much higher than "Private room" listings, so the null hypothesis is not true.

Hypothesis 3: Neighborhood Pricing Disparities

```
> # Calculate the p-value
> p_value <- 2 * pnorm(-abs(z_score))
>
> p_value
[1] 4.839882e-204
```

Our third theory looked at the difference between Manhattan and Brooklyn, specifically the idea that the average price of an Airbnb listing in Manhattan is higher than one in Brooklyn. Because there were so many listings in each neighborhoods group, a z-test was used again.

The z-score showed a big difference in the means, and the p-value that came from it was incredibly small (4.839882e-204), which strongly suggests that there is a big difference in prices between the two neighborhoods. The statistical evidence is so strong that we can say without a doubt that Manhattan listings are, on average, more expensive than Brooklyn listings.

Some Facts About the Analysis

The hypothesis testing phase of the Airbnb dataset analysis has revealed several intriguing facts about the New York City rental landscape on the platform:

1. **No Price Segmentation by Top Housing Market Tier**: As we might expect, Airbnb listings in neighborhoods with home prices in the top 20 percent of the market did not have significantly higher rental prices than listings in other neighborhoods. So, it looks like the higher prices in the housing market don't always show up directly in the prices for short-term rentals on Airbnb. This might be because of different things, like what Airbnb guests want, the kinds of listings that are available, or even the rules that apply in different areas.
2. **Substantial Premium for Entire Homes and Apartments**: The research found that "Entire home/apt" listings are much more expensive than "Private room" listings to rent. The z-p-value tests was almost zero, which means there was a very strong statistical

significance. This fits with what we might expect: whole properties, which have more space and privacy, tend to fetch higher prices. It also shows how popular these kinds of accommodations are, which could be because travelers want more home-like amenities or need to travel with a group.

3. **Manhattan's Significant Price Difference**: A z-test that looked at Airbnb prices in Manhattan and Brooklyn showed that there was a huge difference, with Manhattan having higher prices. It was almost impossible for the p-value to be zero, which means the finding was very strong. This shows that Manhattan is a popular place to stay for Airbnb guests in New York City. This is probably because it is centrally located, close to major attractions, and a cultural and financial hub.
4. **Room Type Over Location**: The big price difference between room types instead of just location is interesting. It suggests that Airbnb guests may care as much about the type of space they rent as they do about the neighborhood it's in. This could be especially helpful for travelers who want to be comfortable, have plenty of space, or "live like a local" during their stay.
5. **Data Robustness Despite Missing Values**: There were a lot of empty values in the reviews per month variable of the Airbnb dataset, but the hypothesis tests were still able to be used confidently. This could mean that other variables, like price, room type, and location, give us enough information to make a good study of the things that affect listing prices.

Bayesian Analysis

Bayesian analysis is a type of statistical analysis that uses what people already know or believe and updates it with new information or evidence. This method works especially well when we don't know what to do, and it's especially useful in markets that are always changing such as Airbnb rentals, where a lot of factors affect price and availability.

Explanation of Bayesian Odds and Its Relevance:

- **Bayesian Odds**: In Bayesian statistics, odds show how likely it is that something will happen compared to how unlikely it is that it will not happen. In particular, the odds version of Bayes' theorem updates odds by combining old odds with new evidence. These new odds are called posterior odds.

- **Relevance**: When it comes to Airbnb prices, Bayesian odds help us figure out how much our guess about how expensive a listing is should change when we get new information, like where the listing is located. This is important for both hosts and guests to know so that hosts can set prices that are competitive, and guests can figure out what the best listings are in different neighborhoods.

Prior Odds Calculation for Expensive Listings:

- **Expensive Listings**: The prior odds of an Airbnb listing being expensive are the ratio of the chance that it is expensive to the chance that it is not expensive. Expensive listings are those that are priced above the 75th percentile of all Airbnb listings.
- **Prior Odds**: The prior odds were found by looking at the percentage of expensive listings. This gave us a general idea of how likely it is that a listing will be expensive before we looked at where it is located.

Likelihood Ratio Calculation Based on Neighborhood Groups:

- **Likelihood Ratio**: That is, this ratio shows how much more likely it is that we will see a listing in Manhattan if we know it is expensive than if we don't know it is expensive.
- **Calculation**: We found this ratio by dividing the chance that a listing is in Manhattan if it's expensive by the chance that it's not expensive.

```
> # Calculate the posterior odds
> posterior_odds_expensive_manhattan <- prior_odds_expensive * likelihood_ratio
>
> posterior_odds_expensive_manhattan
[1] 0.6181832
```

Posterior Odds Calculation and Interpretation:

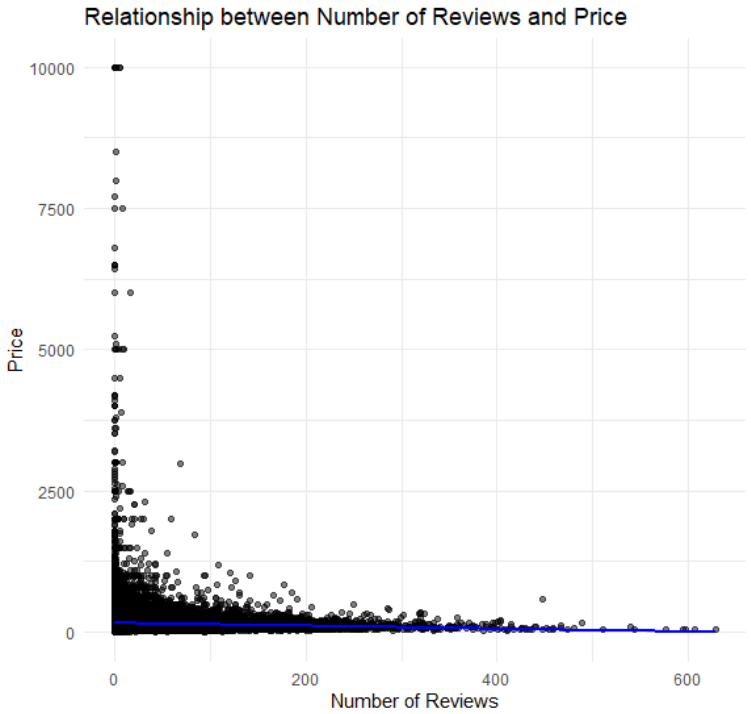
- **Posterior Odds**: The posterior odds were found by multiplying the prior odds by the likelihood ratio. These were the odds that a listing would be expensive after because it was in Manhattan.
- **Interpretation**: After adding the information about location to what we already thought, the posterior odds of 0.618 show that a Manhattan listing is less likely than we first thought to be priced above the 75th percentile. In other words, being in Manhattan does make it more likely that a listing will be expensive, but not as much as one might think based on the odds alone.

This Bayesian analysis shows how location and price on Airbnb are connected in a complex way. The prestige of Manhattan makes us think that listing prices would be higher, but the data shows that this effect isn't as strong as we might have thought before the Bayesian update. In the competitive short-term rental market in New York City, this information is especially helpful for understanding what the market wants and how to set prices.

Data Visualization and Analysis

The visual analysis of the New York City Airbnb dataset has brought forward several interesting relationships and patterns within the data:

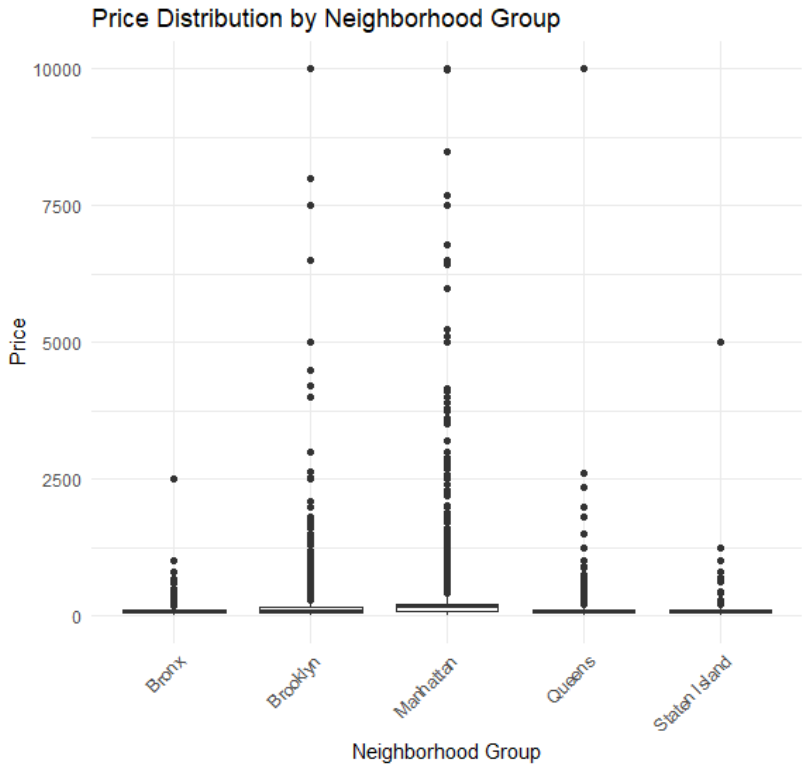
1. Relationship Between Number of Reviews and Price:



There is a wide range of prices for listings with different numbers of reviews, as shown by the scatter plot that shows the relationship between the number of reviews and price. It's interesting

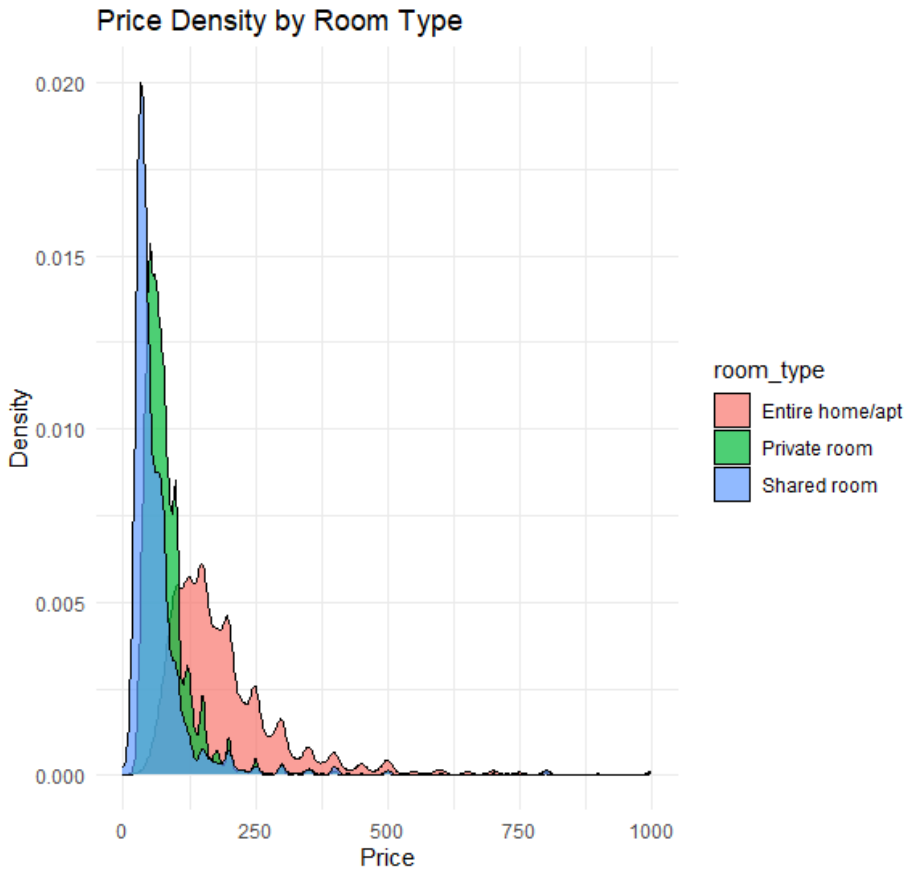
that there doesn't seem to be a strong link between the price of a listing and the number of reviews it has. The plot instead shows that listings with fewer reviews are spread out across all price ranges, while listings with a lot of reviews are more likely to be in the lower to mid-price range. This could mean that listings that are cheaper get booked more often, which leads to more reviews over time.

2. Price Distribution by Neighborhood Group:



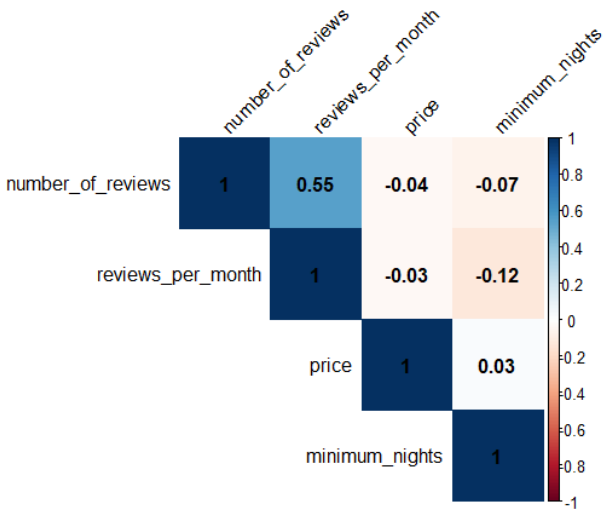
The box plot that shows how prices are spread out among neighborhood groups makes it easy to see how prices vary in New York City's boroughs. Manhattan is the most expensive borough because it has the highest median price and the widest range of prices, with several very expensive homes. Brooklyn also has a wide range of prices, but the median price there is usually not too high or too low. There are fewer Airbnb listings spread out across the Bronx, Queens, and Staten Island, which suggests that the prices are lower.

3. Price Density by Room Type:



It's easy to understand the density plot for price by room type. It shows that whole homes and apartments have a longer tail going up in price, which means they are more likely to be listed at higher price points. The density is higher around the median price. Private rooms have a price peak at lower prices, which means we can usually find them at lower prices. Shared rooms are the least common type of room, and the density plot shows that they are mostly found at the lowest prices.

4. Correlation Heatmap



The correlation heatmap shows how strongly different numerical variables are linked by using colors to show this. The plot shows that there is a moderately positive relationship between the number of reviews and the number of reviews per month. This is to be expected, since more reviews mean a higher overall count. There are, however, only weak links between price and the other variables. This suggests that pricing may be affected by things that aren't included in the dataset.

Conclusion

The journey of analysis has given us a lot of information about the Airbnb market in New York City. The main results show that location, room type, and market dynamics all affect each other in complicated ways. These results go against some common beliefs while supporting other expected trends. This analysis means that both hosts and guests can use data-driven strategies for pricing and choosing accommodations on the Airbnb platform to make it easier to use.

Key Statistical Findings:

- There isn't a statistically significant difference between neighborhoods where home prices are in the top 20 percent and those that aren't when it comes to Airbnb listing prices.
- "Private room" listings are priced a lot less than "Entire home/apt" listings, and the difference is statistically significant.
- It is a strong and statistically significant fact that listings in Manhattan are priced much higher than those in Brooklyn.

Visual Representation of Relationships:

- There wasn't a strong link between the number of reviews and listing prices, as shown by scatter plots.
- Boxplots showed that prices were spread out very differently across different neighborhood groups, with prices being highest in Manhattan.
- Density plots showed that whole homes and apartments were more common and cost more than other types of rooms.

Comparison of EDA Insights with Hypothesis Testing Results:

- The EDA pointed out possible trends and outliers, and hypothesis testing used numbers to show whether these first observations were true or false.
- The EDA showed that there was no link between the number of reviews and price, which matched the results of the permutation test for Hypothesis 1.

Bayesian Analysis Outcomes:

- Bayesian analysis showed that the chances of an Airbnb listing in Manhattan being priced above the 75th percentile are not as high as one might think. This suggests that location is not the only thing that affects pricing.

Discussion

When looking at these results in the context of New York City's rental and housing market, they suggest that location is still important, but the type of accommodation is just as important, if not more so, in setting prices.

Implications for Hosts and Guests:

- Hosts might want to rethink how important location is, especially in high-value areas, and instead focus on making their listings stand out through quality and amenities.
- Guests could find better value by looking at listings with more reviews or considering room types that traditionally offer more competitive pricing.

Limitations and Potential Biases:

- The study might be limited by the number of variables in the datasets; things like recent renovations, the look of the listings, and how responsive the hosts were not considered.
- It's possible that there were biases because listings without reviews were left out or because the data didn't show all neighborhoods equally.

Recommendations

For Hosts:

- Do market research to find out who the competitors are and set the prices of listings accordingly.
- The host might want to spend money on high-quality upgrades and extras that can help them charge more.

For Guests:

- To find value, look beyond location. For example, look at listings with more reviews or choose private rooms.
- Keep in mind that booking patterns and seasonal changes can change prices.

Work Cited

https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data

https://www.kaggle.com/datasets/nelgiriyewithana/new-york-housing-market

https://www.researchgate.net/publication/326814900_Regulating_Airbnb_how_cities_deal_with_perceived_negative_externalities_of_short-term_rentals

https://therealnews.com/ride-along-montreal-launches-first-airbnb-enforcement-squad-in-north-america

https://comptroller.nyc.gov/reports/the-impact-of-airbnb-on-nyc-rents

https://www.researchgate.net/publication/335427185_Online_rental_housing_market_representation_and_the_digital_reproduction_of_urban_inequality