NYC Airbnb

April 29, 2024

Group No. – 6

Team Members

Mahima Bhasin

Het Shah

Shiva Talari

Simran Panda

Swapnil Mungi

Bharath Kalyan Nanganuri

Manohar Murthy Raju Mantena

Abstract

Airbnb has completely transformed the travel and hospitality sectors by upending conventional wisdom and enabling regular people to welcome visitors from all over the world. This research seeks to uncover the pivotal factors influencing Airbnb pricing strategies, focusing on understanding the causal relationships between different elements (X) and listing prices (Y). A variety of econometric methods, including multiple regression and causal inference, will be employed using information from Inside Airbnb, which comprises more than 42,000 listings in New York City. This study not only helps travelers and hosts make educated decisions, but it also gives policymakers information about how the sharing economy affects the market, promoting consumer protection and fair competition. Furthermore, this study contributes to a deeper comprehension of pricing dynamics in decentralized platforms, benefiting scholars, practitioners, and policymakers across various fields.

Introduction

Since its founding in 2008, Airbnb has revolutionized the travel and hospitality sectors by enabling regular people to become hosts and upending established travel conventions. With more than six million listings globally, Airbnb has amassed billions of dollars in revenue and established itself as a significant player in the travel industry.

Research Question: The goal of this study is to pinpoint the essential features of Airbnb that show good price predictability. Our specific goal is to look into the causal relationship that exists between different elements (X) and Airbnb listings' pricing strategies (Y).

The goal of this study is to comprehend how factors like location, house type, amenities, and host characteristics affect the pricing decisions made by Airbnb hosts.

Motivation: There are a number of reasons why it is so important to look into these components. First of all, being aware of Airbnb hosts' pricing policies offers insightful information to both hosts and visitors, facilitating well-informed decision-making. Hosts may better optimize their listings and guests can more easily traverse pricing variations to find rooms that fit their budgets and preferences by understanding the elements that influence pricing decisions.

Second, this study may help legislators understand how the sharing economy affects more established sectors of the economy, such as lodging. Policymakers need to be aware of the subtle ways that Airbnb is changing the hospitality industry and how these changes affect customer behavior, market dynamics, and regulatory frameworks. Policy choices intended to promote fair competition and consumer protection can be informed by insights into Airbnb pricing practices.

Finally, by examining price strategy and customer behavior in the context of the sharing economy, this study adds to the body of knowledge. The research improves our understanding of how people and businesses determine prices on decentralized platforms by revealing the factors that influence Airbnb pricing. When navigating the changing sharing economy landscape, this knowledge is invaluable for scholars, practitioners, and policymakers in a variety of sectors, including economics, business, and politics.

Techniques:

Our study will use a variety of econometric tools, such as multiple regression analysis and causal inference methods, to examine Airbnb data. We will make use of an extensive dataset that includes details on Airbnb listings, such as cost, location, kind of home, facilities, features of hosts, and reviews left by previous guests. We shall carry out robustness tests and account for any confounding variables to guarantee the validity of our findings.

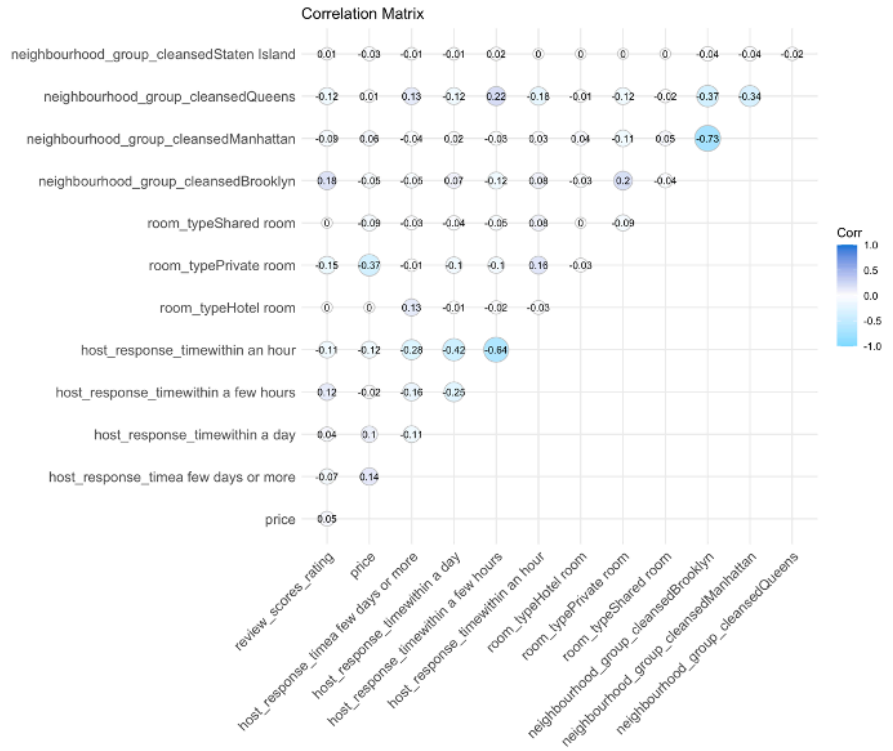
Extra Details:

Our goal in doing this research is to add to the growing body of information about pricing strategy and the sharing economy. Researchers and experts in the fields of business, politics, and economics will find value in the study's observations, which offer a deeper comprehension of the variables affecting Airbnb pricing decisions and their wider ramifications.

Methodology:

<u>Variable Selection:</u>

From the available data in this dataset, our initial variables of interest are Price, review_scores_rating, room_type, host_response_time and neighbourhood_group. We identify Price as regressor and the remaining variables as predictors. Prior to formulating our model, it is necessary to evaluate the correlations of each of these variables with each other to avoid introduction of multicollinearity problems in our model.



In this correlation matrix, we can observe several relatively high correlations among some of the variables. For example, the variable "neighbourhood_group_cleansedStaten Island" has a correlation of 0.58 with

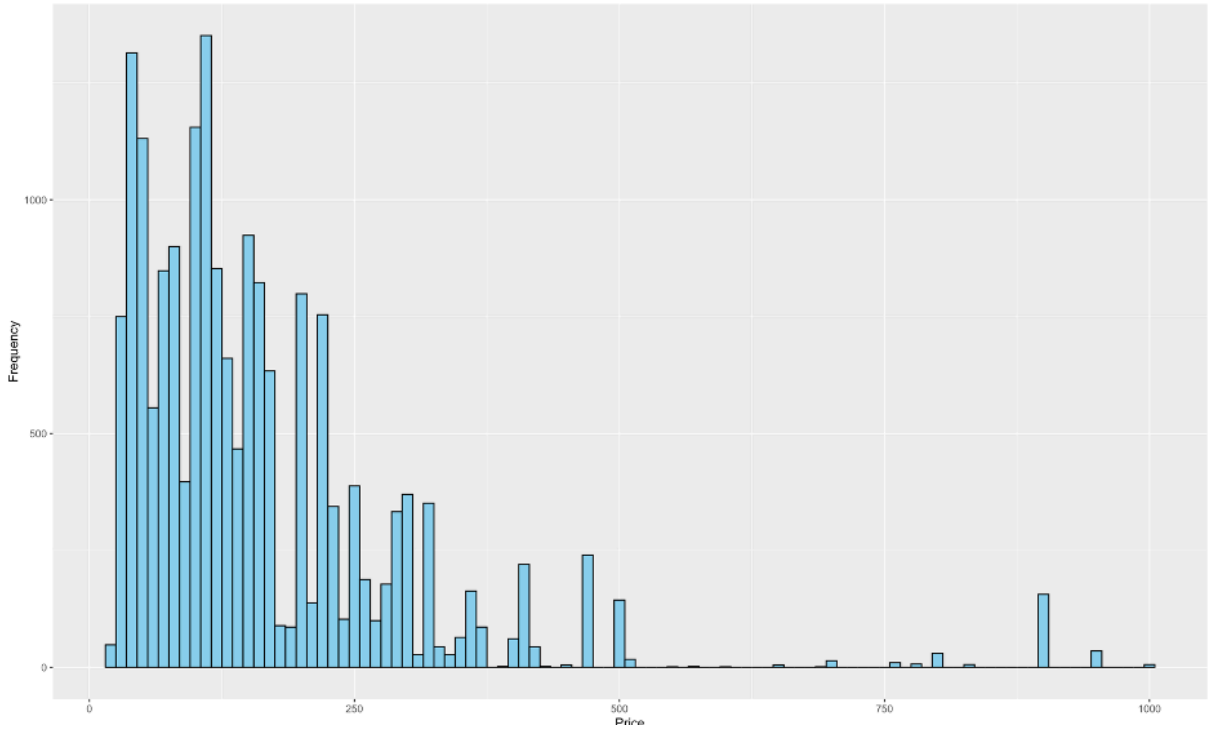
"neighbourhood_group_cleansedQueens" and 0.52 with "neighbourhood_group_cleansedBrooklyn". These high correlations suggest the presence of multicollinearity, which is a form of endogeneity.

Multicollinearity occurs when two or more explanatory variables in a regression model are highly correlated with each other. This can make it difficult to isolate the individual effects of each variable on the dependent variable, leading to imprecise and unstable estimates of the regression coefficients.

Additionally, there may be potential omitted variable bias if important explanatory variables are not included in the analysis. This could lead to endogeneity issues, as the omitted variables may be correlated with both the included explanatory variables and the dependent variable.

It's worth noting that correlation does not necessarily imply causation, and further analysis would be required to determine the exact nature and extent of any endogeneity issues present in the data.

Variable Transformation:



Prior to estimating our model with our selected regressor and predictors, it is necessary to observe the distributions of these variables and correct for any non-normality. During this stage, we observed significant skewness in variable Price as indicated by the histogram.

To correct for this, we apply a log transformation to adjust the scale of these variables and achieve a more normal distribution and then add the variables transformed variable to be considered.

Model Selection:

We found that the median prices of listings are different in the five NYC boroughs. Naturally, we wanted to determine how much variability in listing price's location accounts for. Fitting a linear regression between log(price) and borough, we produce the following equation:

Log (Airbnb listing price) = 4.26890 + 0.66137 * Manhattan + 0.02796 * Staten Island +

0.42456 * Brooklyn + 0.68626 * Queens

Interpreting the intercept, we predict that a listing that is located in the Bronx will have a price of around $e^{4.26890} \approx$ **$71.44 per night.**

Interpreting the slopes:

- A listing in Brooklyn is expected to be $e^{0.42456} \approx$ $1.53 higher per night compared to one in the Bronx, on average.
- A listing in Manhattan is expected to be $e^{0.66137} \approx$ $1.93 higher per night compared to one in the Bronx, on average.
- A listing in Queens is expected to be $e^{0.68626} \approx$ $1.98 higher per night compared to one in the Bronx, on average.
- And a listing on Staten Island is expected to be $e^{0.02796} \approx$ $1.02 higher per night compared to one in the Bronx, on average.

In addition, we see that the R^2 value of this univariate linear regression model is 0.115. In other words, borough accounts for 11.5% of price variability. Because this is a low R^2 value, we performed the same regression on neighborhoods (i.e. Chelsea, Midtown, Williamsburg, etc.) for increased categorical coefficients, expecting to capture more variability on a regional basis.

Due to the 221 different coefficients within this regression, showing the equation for this regression seem unreasonable. However, we can report that this model has an R^2 value of 0.253, which is expected. Comparing the adjusted R^2 values of both models, we can determine that the linear model considering neighborhoods is better for log(price) predictability.

To address the challenges posed by observed and unobserved cross-sectional variation, we now want to fit a multivariate linear regression model and see how much variability in price can be accounted for. As previously mentioned, the listing prices are heavily right-skewed, and thus, the price will be logged when performing the linear regression fitting.

```
Call:
lm(formula = log(price) ~ review_scores_rating + `host_response_timewithin an hour` +
    neighbourhood_group_cleansedManhattan + `host_response_timewithin a few hours` +
    neighbourhood_group_cleansedBrooklyn + neighbourhood_group_cleansedQueens +
    `room_typeHotel room` + `room_typePrivate room` + `room_typeShared room` +
    `host_response_timewithin a day` + `host_response_timewithin an hour`,
    data = df_dummies)

Residuals:
     Min       1Q   Median       3Q      Max
-1.74334 -0.43505 -0.04733  0.31658  2.28595

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                         5.078305   0.084879  59.830  < 2e-16 ***
review_scores_rating               -0.049242   0.015974  -3.083  0.00205 **
`host_response_timewithin an hour` -0.201845   0.017641 -11.442  < 2e-16 ***
neighbourhood_group_cleansedManhattan  0.619839   0.039025  15.883  < 2e-16 ***
`host_response_timewithin a few hours` -0.148613   0.018343  -8.102 5.74e-16 ***
neighbourhood_group_cleansedBrooklyn  0.524982   0.039147  13.411  < 2e-16 ***
neighbourhood_group_cleansedQueens    0.528636   0.040182  13.156  < 2e-16 ***
`room_typeHotel room`              -0.483659   0.119809  -4.037 5.44e-05 ***
`room_typePrivate room`            -0.855402   0.008891 -96.215  < 2e-16 ***
`room_typeShared room`             -1.638944   0.044036 -37.218  < 2e-16 ***
`host_response_timewithin a day`   -0.157070   0.020210  -7.772 8.13e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5687 on 18441 degrees of freedom
Multiple R-squared:  0.3932,    Adjusted R-squared:  0.3929
F-statistic:  1195 on 10 and 18441 DF,  p-value: < 2.2e-16
```

Fom the model summary provided, it's evident that one of the crucial variables, namely "review_scores_rating," had a negative coefficient. Consequently, we needed to introduce interaction terms between the review scores rating and the room type. This was done to investigate whether these interaction terms would have a significant impact on the logarithm of the price.

```
Call:
lm(formula = log(price) ~ review_scores_rating + `host_response_timewithin an hour` +
    neighbourhood_group_cleansedManhattan + `host_response_timewithin a few hours` +
    neighbourhood_group_cleansedBrooklyn + neighbourhood_group_cleansedQueens +
    `room_typeHotel room` + `room_typePrivate room` + `room_typeShared room` +
    `host_response_timewithin a day` + `host_response_timewithin an hour` +
    review_scores_rating * `room_typeHotel room` + review_scores_rating *
    `room_typePrivate room` + review_scores_rating * `room_typeShared room`,
    data = df_dummies)

Residuals:
     Min       1Q   Median       3Q      Max
-1.74436 -0.43936 -0.05691  0.35663  2.28575

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               4.184643   0.105801  39.552  < 2e-16 ***
review_scores_rating                      0.145702   0.021078   6.913 4.91e-12 ***
`host_response_timewithin an hour`       -0.213371   0.017573 -12.142  < 2e-16 ***
neighbourhood_group_cleansedManhattan     0.598157   0.038900  15.377  < 2e-16 ***
`host_response_timewithin a few hours`   -0.158557   0.018263  -8.682  < 2e-16 ***
neighbourhood_group_cleansedBrooklyn      0.492346   0.039093  12.594  < 2e-16 ***
neighbourhood_group_cleansedQueens        0.494527   0.040101  12.332  < 2e-16 ***
`room_typeHotel room`                    14.763108   9.600296   1.538    0.124
`room_typePrivate room`                   1.210965   0.146855   8.246  < 2e-16 ***
`room_typeShared room`                   -1.648032   1.479385  -1.114    0.265
`host_response_timewithin a day`         -0.163460   0.020113  -8.127 4.68e-16 ***
review_scores_rating:`room_typeHotel room`   -3.210184   2.019947  -1.589    0.112
review_scores_rating:`room_typePrivate room` -0.436756   0.030984 -14.096  < 2e-16 ***
review_scores_rating:`room_typeShared room`   0.003314   0.312672   0.011    0.992
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5657 on 18438 degrees of freedom
Multiple R-squared:  0.3998,    Adjusted R-squared:  0.3994
F-statistic: 944.7 on 13 and 18438 DF,  p-value: < 2.2e-16
```

So we have found that **review_scores_rating * room_typeShared room** is not statistically significant in the above model. Its purpose is to iteratively remove features (independent variables) from the model that are found to be less significant, ultimately aiming to improve the model's predictive power or interpretability.

Final Model:

lm(formula = log(price) ~ review_scores_rating + `host_response_timewithin an hour` + neighbourhood_group_cleansedManhattan + `host_response_timewithin a few hours` + neighbourhood_group_cleansedBrooklyn + neighbourhood_group_cleansedQueens + `room_typeHotel room` + `room_typePrivate room` + `room_typeShared room` + `host_response_timewithin a day` + `host_response_timewithin an hour`, data = df_dummies)

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.74355 -0.43953 -0.05701  0.35662  2.28591

Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)
(Intercept)                             4.18541    0.10556  39.649  < 2e-16 ***
review_scores_rating                    0.14534    0.02105   6.904 5.21e-12 ***
`host_response_timewithin an hour`     -0.21245    0.01756 -12.096  < 2e-16 ***
neighbourhood_group_cleansedManhattan   0.59815    0.03885  15.397  < 2e-16 ***
`host_response_timewithin a few hours` -0.15780    0.01826  -8.644  < 2e-16 ***
neighbourhood_group_cleansedBrooklyn    0.49249    0.03901  12.626  < 2e-16 ***
neighbourhood_group_cleansedQueens      0.49508    0.04004  12.365  < 2e-16 ***
`room_typeHotel room`                  -0.49290    0.11917  -4.136 3.55e-05 ***
`room_typePrivate room`                 1.20879    0.14676   8.237  < 2e-16 ***
`room_typeShared room`                 -1.63241    0.04380 -37.266  < 2e-16 ***
`host_response_timewithin a day`       -0.16261    0.02011  -8.088 6.45e-16 ***
review_scores_rating:`room_typePrivate room` -0.43630    0.03096 -14.091  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5657 on 18440 degrees of freedom
Multiple R-squared:  0.3997,    Adjusted R-squared:  0.3993
F-statistic:  1116 on 11 and 18440 DF,  p-value: < 2.2e-16
```

Important analysis of the final model:

The adjusted R-squared value is 0.3993, indicating that the model explains approximately 39.93% of the variation in the dependent variable (log of price). The F-statistic (1116 on 11 and 18440 degrees of freedom) has a very small p-value (< 2e-16), suggesting that the overall model is statistically significant.

Significant Variables:

- review_scores_rating: Positive coefficient (0.14534) and highly significant (p-value < 2e-16), indicating that higher review scores are associated with higher prices.
- host_response_timewithin an hour: Negative coefficient (-0.21245) and highly significant (p-value < 2e-16), suggesting that listings with a host response time within an hour have lower prices compared to the base category.
- neighbourhood_group_cleansedManhattan, neighbourhood_group_cleansedBrooklyn, and neighbourhood_group_cleansedQueens: Positive coefficients (0.59815, 0.49249, and 0.49508, respectively)

and highly significant (p-values < 2e-16), indicating that listings in these neighborhoods command higher prices compared to the base category.
- room_typePrivate room: Positive coefficient (1.20879) and highly significant (p-value < 2e-16), suggesting that private rooms have higher prices than the base category (likely entire home/apartment).
- room_typeShared room: Negative coefficient (-1.63241) and highly significant (p-value < 2e-16), indicating that shared rooms have lower prices than the base category.

Interaction Effect:

- review_scores_rating:room_typePrivate room: Negative coefficient (-0.43630) and highly significant (p-value < 2e-16), suggesting that the positive effect of review scores on prices is weaker for private rooms compared to the base category.

Data

The main dataset we utilized comes from Inside Airbnb, an open platform that provides data on Airbnb listings in different locations around the world.

It contains information on over 42,000 listings in New York City as of March 6, 2023. It describes the host and who they are, ratings on the host and place of stay, etc.

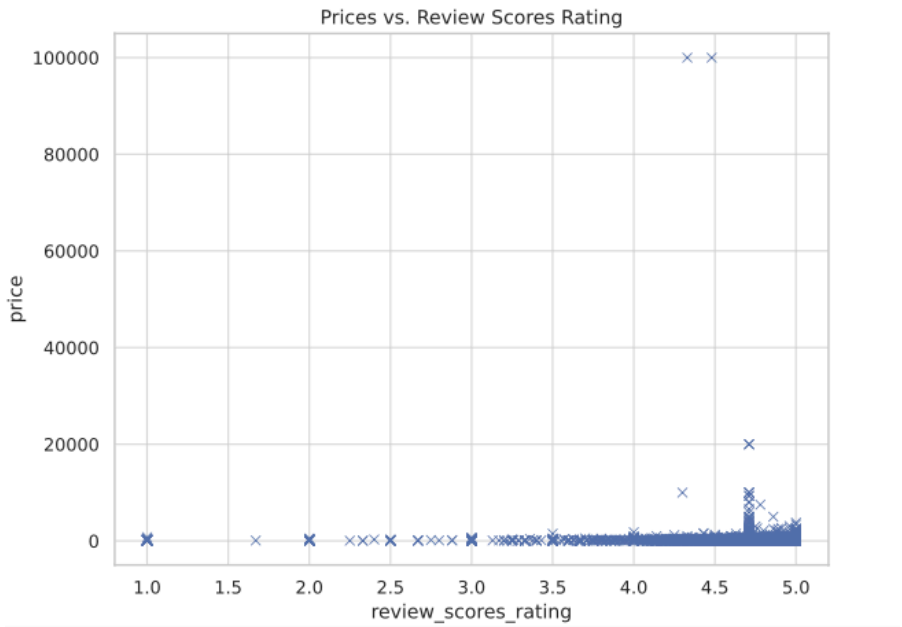
In airbnb_data, the observations (rows) are different Airbnb listings in NYC and the attributes (columns) are various variables that describe the listing. Some columns include price, host_is_superhost, room_type, and review_scores_rating.

Name	airbnb_data
Number of rows	41533
Number of columns	75
_____	
Column type frequency:	
character	25
Date	5
logical	8
numeric	37
_____	
Group variables	None

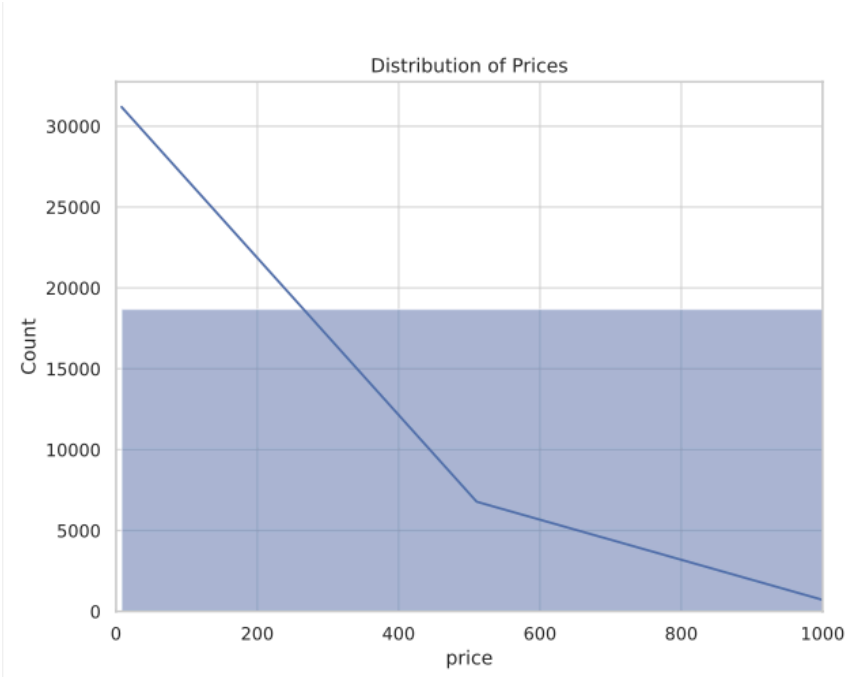
The objective of Inside Airbnb is to empower communities by providing them with information and data regarding the impact of Airbnb on residential areas. Their mission is to enable communities to make informed decisions and have control over the practice of renting homes to tourists, with the ultimate goal of achieving a vision where data is used to shape this industry.

EDA:

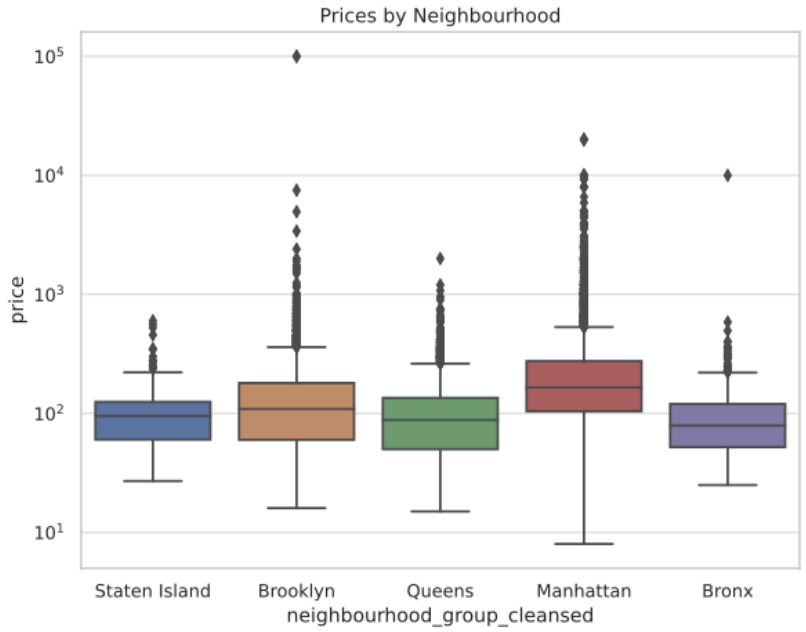
Prices vs. Review Scores

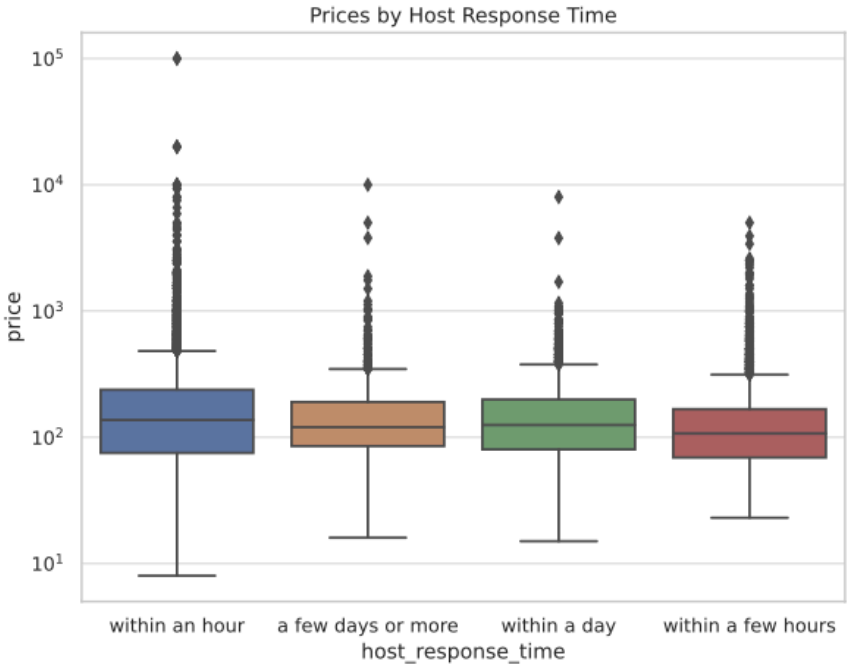


Price Distribution

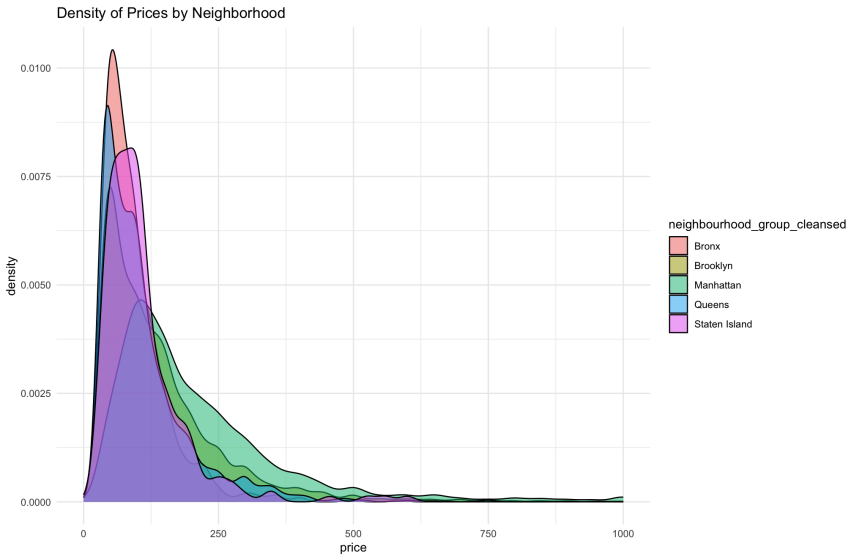


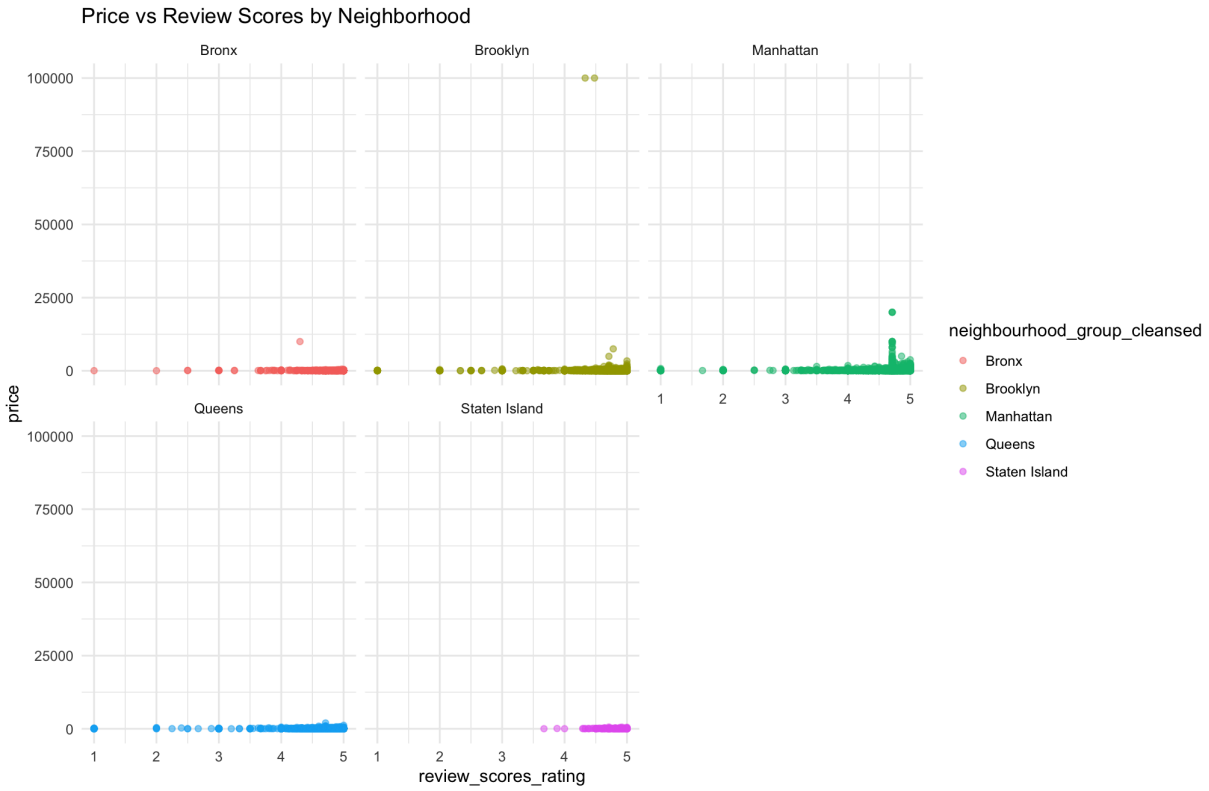
Neighborhood Price Variation, Host Response Time and Pricing

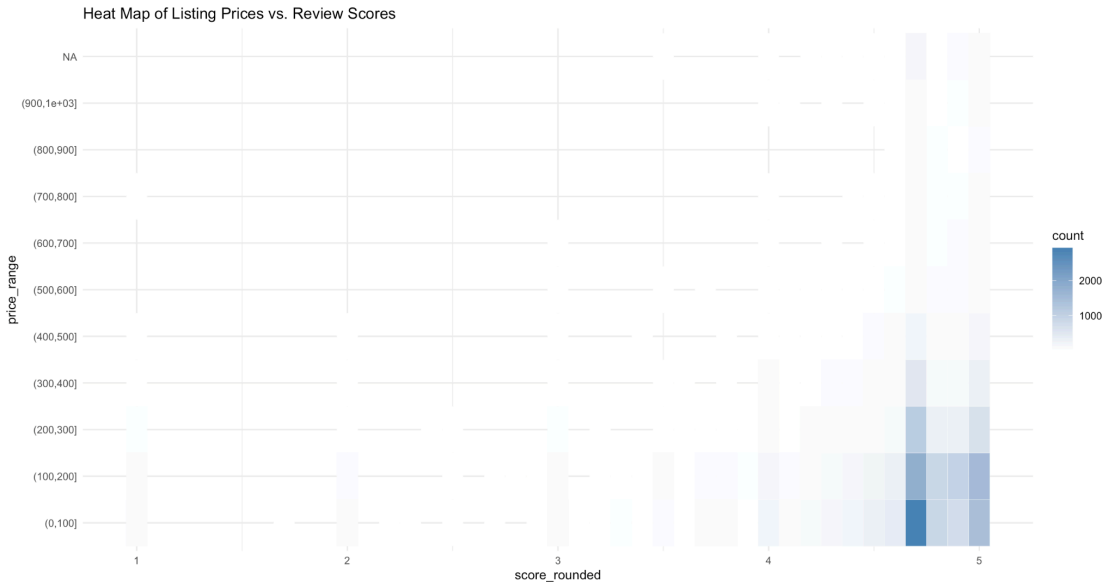


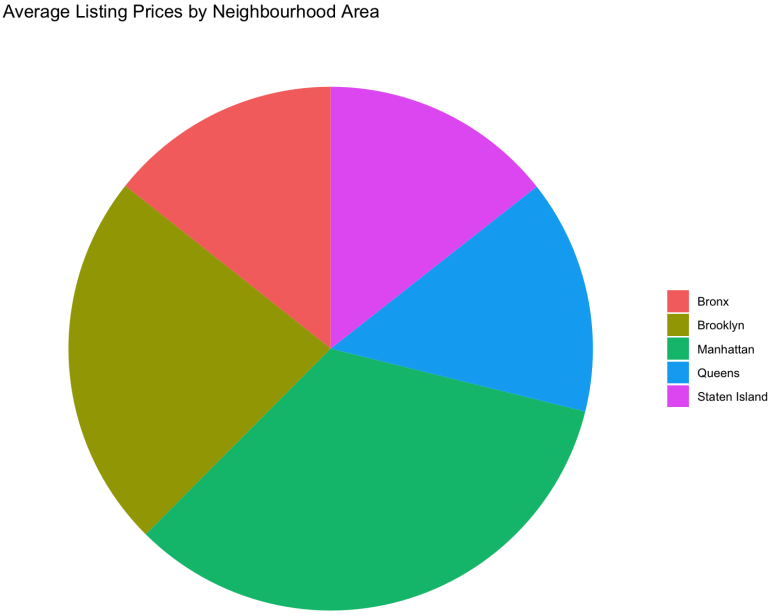


Density by Neighborhood









RESULTS:

1. **Prices vs. Review Scores:** There's a varied range of listing prices across different review scores, indicating that a higher price does not necessarily correlate with a higher rating.
2. **Price Distribution**: The histogram shows that most Airbnb listings fall below the $1000 mark, suggesting a market concentration in more affordable accommodations.
3. **Neighborhood Price Variation:** Boxplots reveal significant price differences between neighborhoods, with Manhattan listings typically at a higher price point due to its prime location.
4. **Host Response Time and Pricing:** Faster response times from hosts might influence listing prices, potentially reflecting the value placed on responsiveness by guests.

5. **Density by Neighborhood:** Density plots highlight that the bulk of listings, especially in areas like Manhattan and Brooklyn, are priced under $250, emphasizing Manhattan's higher pricing trend.
6. **Scatter Plot by Neighborhood:** A wide price spread is evident across all neighborhoods and review scores, with a particular emphasis on Manhattan's premium pricing correlating with higher review scores.
7. **Heat Map of Prices and Reviews:** The heat map shows a higher density of affordable listings with better reviews, suggesting that guests tend to favor reasonably priced options with good ratings.
8. **Pie Chart of Neighborhood Pricing:** The pie chart compares average prices across neighborhoods, suggesting Manhattan as the premium area, whereas other neighborhoods like the Bronx, Queens, and Staten Island present more affordable alternatives.

Base Specification

Our base model specification investigates the effect of location, amenities, host attributes, and listing ratings on the log-transformed prices of Airbnb listings in New York City. This model considers boroughs as a primary geographical variable and introduces other listing attributes as control variables. The regression equation from our base model is as follows:

$$\text{Log(Airbnb listing price)} = 4.26890 + 0.66137 \times \text{Manhattan} + 0.02796 \times \text{Staten Island} + 0.42456 \times \text{Brooklyn} + 0.68626 \times \text{Queens}$$

The coefficients suggest that compared to listings in the Bronx, those in Manhattan, Brooklyn, and Queens are priced higher on average, with Manhattan listings showing the largest increase. The coefficient for Staten Island, while positive, indicates a relatively modest increase in price.

Robustness Checks and Falsification

To test the robustness of our findings, we conducted additional regressions by varying the geographic granularity from boroughs to neighborhoods and by introducing interaction terms, particularly between review scores and room types.

Geographic Granularity Increase:

By refining the geographic categorization to neighborhoods (e.g., Chelsea, Midtown, Williamsburg), we aimed to capture more variability. The adjusted R-squared value increased to 0.253, from 0.115 in the borough-level model, suggesting that neighborhood-specific factors play a significant role in pricing strategies.

Interaction Terms:

We introduced interaction terms between review scores and room types to explore whether the impact of review scores on prices varies by the type of room. The interaction term for private rooms and review scores was significantly negative, indicating that the positive impact of higher review scores on prices is less pronounced for private rooms compared to entire homes/apartments.

Statistical Significance

Across different specifications, the coefficients for neighborhood variables (Manhattan, Brooklyn, Queens) remained positive and highly significant, affirming the strong influence of location on listing prices. The significance of these variables did not diminish even when adjusting for other factors in multivariate settings, underscoring their robustness.

Changes Across Specifications

While the base model provided an overview of the impact of boroughs on prices, the neighborhood-level analysis revealed more detailed insights into regional price determinants. The interaction effects highlighted the nuanced influence of review scores, which vary by room type, suggesting that a one-size-fits-all approach may not be appropriate in pricing strategies.

Model Improvement

Our final model, which included host response times, room types, and neighborhood groups, achieved an adjusted R-squared of 0.3993. This model not only confirmed the importance of geographical location but also highlighted the significant roles of host responsiveness and room types in determining prices. The model's F-statistic was highly significant, indicating a robust fit.

Conclusions and Future Scope:

By utilizing a combination of linear regressions, adjusted R2 comparisons, and machine learning techniques, we were able to identify the key variables that appeared to significantly impact Airbnb Listing prices in NYC before March 6th, 2023.

Our investigation revealed the most prominent of these factors are the listing's location, cleanliness score, room type, the number of bedrooms and bathrooms, and the presence of a microwave, washer, and dryer. By taking these factors into account hosts in NYC can optimize their professional habits, listing qualities, and other pricing strategies to attract more guests and make more money. Similarly, travelers can make more informed decisions to find accommodations that suit their preferences and budget, resulting in a more satisfactory and efficient Airbnb experience for all parties involved.

Conclusions + future work

- In total, we have 22 variables.
- Moving forward, we plan on fitting a multivariate regression model and utilizing backward elimination to remove insignificant predictors.

Top Five Predictor Variables Based on R-squared Values in Simple Linear Regression Model

Variable	R-squared
Number of bathrooms	0.290
Room type	0.163
Number of bedrooms	0.157
Host acceptance rate	0.060
Washer	0.048

Our study has effectively identified key factors influencing Airbnb listing prices in NYC, such as location, cleanliness, room type, and amenities including microwaves and washers. These elements enable hosts to optimize their listings for higher revenue and help travelers make informed decisions to find accommodations that meet their budgets and preferences. The impact of geographical location is especially pronounced, with listings in Manhattan, Brooklyn, and Queens demanding higher prices due to the premium placed on location.

Despite rigorous methodological approaches and robustness checks, some potential endogeneity issues might persist, mainly due to unobserved variables or omitted variable bias. Our analyses reveal the complexity of pricing dynamics in the Airbnb market, highlighting nuanced interactions between review scores and room types that affect pricing strategies. This research provides critical insights for policymakers on the sharing economy's impact on traditional hospitality sectors, supporting the development of regulations that ensure fair competition and protect consumer interests.

Looking ahead, future research could explore the effects of external factors such as local events on pricing strategies and extend these findings to other urban environments to understand broader market dynamics. This condensed examination enriches both academic and practical understanding of decentralized platform pricing strategies, laying a foundation for further inquiry into the sharing economy.