# How Neighborhood, Host, and Listing Characteristics Affect the Price of Airbnb Listings in New York City

William Afework*

August 24, 2024

**Abstract**

This paper investigates the factors affecting Airbnb listing prices in New York City, employing datasets comprising neighborhood characteristics, host details, tourism attraction, census data and listing specifics. The analysis makes use of OLS regressions and Random Forest Model's, as well as an original approach for measuring tourism activity, revealing that proximity to popular tourist attractions exerts the most significant influence on pricing, overshadowing traditional supply metrics like available housing units. Contrary to expectations, listings with more active reviews and higher minimum housing units correlate with lower listing prices, suggesting that suggesting that high availability and frequent turnover might incentivize hosts to competitively price their listings to attract more bookings. Furthermore, room type emerges as a strong price determinant, with entire homes and apartments commanding price premiums over private rooms. Interestingly, the study also revealed the number of listings a host owns distinctly impacted their pricing strategy. The study contributes to the burgeoning literature on the sharing economy's impact on urban housing markets, providing insights into the complex pricing dynamics within one of the world's most iconic cities.

*University of Toronto

# 1 Introduction

With the rapid rise of Airbnb, it is important for stakeholders to understand the factors affecting pricing of listings. Such research has numerous applications: In improving the efficiency of Airbnb and short term rental markets; for policy makers to better understand and develop regulations around the market, and for other participants such as tourist services and investors to better understand the platform. New York City, a major financial center and economic powerhouse with a bustling tourism sector, diverse population, rich culture presents an interesting case study for our question.

Analyzing varying neighborhood characteristics, such as location, safety, landmarks, and local amenities, may uncover their role in shaping Airbnb prices. These factors not only affect the desirability and demand for listings but also reflect the socio-economic status and demographic composition of neighborhoods. Additionally, listing characteristics, including property type, size, amenities, reviews, directly influences the prices chosen by hosts. This research posits that a detailed analysis of these characteristics will provide insights into the economic principles governing the sharing economy's impact on urban housing markets.

Previous studies into this line of research have uncovered interesting results. A study examining Airbnb prices in Bristol found that the room and property type were the biggest drivers of price, and that while 'house' type properties were more expensive the closer they were to the city centre, non house type properties commanded a premium the further away they were (Voltes-Dorta and Sánchez-Medina, 2020). Another study examining Metro Nashville found that the distance to landsmarks, conventions centres and highways were positive drivers in price, but also unusually found a significant negative correlation between the number of reviews and rating with listing price (Zhang et al., 2017). Suggesting that a higher number of reviews and rating correlated with a lower price. Another study found that an Airbnb's proximity to tourist areas had a postive correlation on price (Perez-Sanchez et al., 2018). One more interesting study focusing on Barcelona also finds that Airbnb listings are concentrated in the city centre and the beach (Gutiérrez et al., 2017).

Another comprehensive study looking into price determinants found that a listing's price depended both on the listing's characteristics and reputation, and the host's involvement (measured by various platform activities) and experience. They found that the host's involvement and listing characteristics were by far the two biggest price determinants. A higher review score and higher distance to the city center also had a negative impact on price, while stricter rental policies resulted in higher prices (Toader et al. 2020). Though other studies have found the opposite and that a higher review score is associated with higher prices (Gibbs et al, 2017). Another study examining both listings features, host characteristics, and neighborhood characteristics found that after controlling for all renter visible factors such as listing features, reviews, and area demographics, that the race of the host was a significant price differential, with Hispanic and Asian hosts charging an average 8-10 percent lower price relative to white hosts (Voelz et al. 2017).

We will contribute to the literature by performing a comprehensive analysis examining how neighborhood, host, and listings characteristics make up the determinants of Airbnb price. In terms of novel additions we will also scrape data for tourist attractions in New York City. This improves upon other methods of examining how location affects price, which are typically limited to calculated straight line distances to only a handful of standardized landmarks such as city Hall's and city centers (Teubner et al., 2017; Wang and Nicolau, 2017).

This research fundamentally answers the economic determinants of Airbnb pricing in New York City by integrating traditional econometric techniques and advanced machine learning models. Using Ordinary Least Squares (OLS) regressions, it quantitatively assesses how various factors, including neighborhood characteristics and listing specifics, impact prices. Complementing this, Random Forest models uncover non-linear interactions and provide a deeper understanding of market dynamics.

# 2 Data

## 2.1 Airbnb Dataset

The primary dataset utilized in this study was sourced from Kaggle, a platform that hosts a wide range of data science projects and competitions. This particular dataset originates from Inside Airbnb, an independent, non-commercial set of data tools that provides data and analytics for Airbnb listings across various global cities. The dataset selected for this analysis is labeled as the "summary" dataset for New York City, reflecting data from the year 2019.

Inside Airbnb's "summary" dataset offers aggregated information and metrics tailored to analyze the performance and characteristics of Airbnb listings within New York City. It includes variables such as listing prices, room types, number of reviews, average review scores, and the geographical coordinates of each listing. Each record in the dataset represents an individual Airbnb listing, making each listing the primary observation level.

The data encompasses a variety of attributes that are critical for understanding the dynamics of the Airbnb market, including but not limited to; price, room type, reviews per month, etc. The dataset contains 48895 listings, however in our regressions we dropped listings that had a minimum nights over 28 days as there was very limited on these listings for some of our categorical variables (certain room types and neighborhoods), resulting in numerical and precision issues. This left us with 44048 listings.

## 2.2 Tourist Attractions Dataset

In addition to the Airbnb listing data, this study incorporates a uniquely constructed dataset of tourist attractions in New York City, which was developed to better understand the influence of tourism proximity on Airbnb pricing. The tourist attractions data was sourced through the Google Places API, a service that provides detailed information about geographic locations and points of interest. The data collection focused on key tourist spots across New York City, capturing attributes such as location coordinates, visitor ratings, and types of attractions.

To quantify the impact of these tourist attractions on Airbnb listings, a 'normalized weighted rating' was created for each attraction, representing its popularity. This rating was calculated by adjusting the raw visitor ratings for the volume of reviews, thereby providing a measure of both the attraction's quality and its appeal. Subsequently, a 'tourism proximity score' was developed for each Airbnb listing. This score was derived by considering the distance of each listing from nearby attractions and weighting these distances by the attractions' normalized ratings. The rationale behind this scoring system was to not only consider how close a listing is to any tourist attraction but also how significant and popular that attraction is.

The resulting 'tourism proximity score' was then merged into the primary Airbnb dataset. This enriched dataset offers a more nuanced understanding of how proximity to popular tourist destinations influences the pricing strategies adopted by Airbnb hosts in New York City. The integration of this data allows for a comprehensive analysis of the spatial economic dynamics at play, providing insights into the premium that proximity to high-value tourist locations can command in the short-term rental market.

## 2.3 Census Dataset

This study also integrates data from the 2020 New York City Department of City Planning census. Although the census offers a wide array of demographic and housing data, this research specifically utilizes the 'vacant housing units' variable. The inclusion of this variable is based on the hypothesis that vacancy rates in a neighborhood can influence Airbnb listing prices, potentially reflecting the availability or supply of housing in the area.

The vacant housing units data provides valuable insights into the housing market dynamics of different neighborhoods. High vacancy rates might indicate a surplus of available properties, possibly leading to competitive pricing strategies among Airbnb hosts in these areas. Conversely, lower vacancy rates could signify a tighter housing market, potentially driving up the prices of available Airbnb listings due to increased demand. By merging this census data with the Airbnb and tourist attractions datasets, the study gains a more robust framework for analyzing how residential market conditions interact with tourism-related factors to shape pricing strategies within New York City's Airbnb market.

## 2.4 American Community Survey Dataset

To provide a deeper understanding of socioeconomic influences on Airbnb pricing, this research incorporates data from the 2019 American Community Survey (ACS), conducted by the U.S. Census Bureau. Specifically, this study utilizes the 'median neighborhood income' variable, which serves as a crucial indicator of the economic status of the neighborhoods where Airbnb listings are located.

Median neighborhood income is a pivotal factor in pricing analysis as it often correlates with consumer spending power and preferences. Higher median incomes in a neighborhood typically suggest that residents and visitors might be willing to pay more for accommodations, reflecting higher demand for quality and services. Conversely, areas with lower median incomes might experience price sensitivity, influencing Airbnb hosts to adjust their pricing strategies accordingly. Integrating this variable allows for a nuanced examination of how economic conditions at the neighborhood level influence the pricing of Airbnb listings, providing insights into the economic stratification and its impact on the sharing economy within urban settings.

By combining the median neighborhood income data with the Airbnb, tourist attractions, and census data, the analysis captures a comprehensive picture of the interplay between economic status, tourism dynamics, and housing availability, contributing to a more informed discussion on market behaviors in New York City's Airbnb sector.

# 3 Visualizations and Summary Statistics

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| reviews_per_month | 38843 | 1.37 | 1.68 | 0.01 | 0.19 | 0.72 | 2.02 | 58.5 |
| minimum_nights | 48895 | 7.03 | 20.51 | 1.00 | 1.00 | 3.00 | 5.00 | 1250.00 |
| number_of_reviews | 48895 | 23.27 | 44.55 | 0.00 | 1.00 | 5.00 | 24.00 | 629.00 |
| calculated_host_listings_count | 48895 | 7.14 | 32.95 | 1.00 | 1.00 | 1.00 | 2.00 | 327.00 |
| price | 48895 | 152.72 | 240.15 | 0.00 | 69.00 | 106.00 | 175.00 | 10000.00 |

Table 1: Summary Statistics of Airbnb Variables

The summary statistics of our selected Airbnb variables (Table 1) reveal a multifaceted portrait of the sharing economy's accommodation sector. The mean reviews per month (1.37) reflect a moderate level of user engagement across listings, with a relatively high standard deviation (1.68) indicating variability in the popularity or utilization of listings. The significant range between the minimum (0.01) and maximum (58.5) reviews per month further underscores the heterogeneity in listing activity, possibly due to differences in location desirability, listing quality, and host responsiveness.

Minimum nights required by listings present an average of 7.03 nights, suggesting a market inclination towards week-long stays. However, the large standard deviation (20.51) and a broad max value (1250) imply that some hosts may be strategically targeting longer-term stays, potentially to reduce turnover costs or to align with regulatory requirements aimed at mitigating the impact on local housing markets.

The average number of reviews per listing stands at 23.27, a testament to the platform's capacity for generating repeat usage and the maturity of listings. The wide gap evidenced by the standard deviation (44.55) and the maximum number of reviews (629) may indicate a subset of highly favored listings, which could be a function of superior service, strategic pricing, or advantageous location, thereby serving as a bellwether for market success.

Notably, the calculated host listings count averages at 7.14, with a high standard deviation (32.95), revealing the presence of both individual hosts and professional operators within the market. The maximum value (327) indicates the emergence of 'superhosts' who manage an extensive portfolio of properties. We can speculate that these hosts who manage a large of properties enjoy economic success, and thus might act differently on the market and incorporate different pricing strategies.

Lastly, the average price of listings is noted at \$152.72, with a pronounced standard deviation (240.15), reflecting a broad spectrum of pricing strategies and property types in the Airbnb market. The substantial range from the minimum (\$0) to the maximum (\$10000) price points to the existence of diverse accommodation options, from budget-friendly rooms to high-end luxury estates, indicating a market that caters to a wide array of consumer preferences and financial capabilities.

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Tourism Proximity | 44048 | 10182 | 2803 | 2798 | 8234 | 9620 | 11804 | 42499 |
| VacHUs | 44048 | 76030 | 19589 | 10490 | 67850 | 67850 | 96144 | 96144 |
| Median Household Income | 44048 | 78444 | 13852 | 41432 | 66937 | 72561 | 93651 | 93651 |

Table 2: Summary Statistics of Merged Features

The analysis of added variables (which will be considered neighborhood variables in our regressions) provides valuable insights into local economic conditions. The proximity to tourist attractions also incorporates a weighed rating of popularity as indicated by google reviews, has an average value of 10,182. The standard deviation of 2,803 in this metric not only underscores the diverse appeal of listings in relation to tourist landmarks but also reflects the variation in their digital footprint and perceived attractiveness. The expansive range from 2,798 to 42,499 indicates a variety of attractions from less frequented, possibly niche, locations to those that are established centers of tourism.

The 'VacHUs' variable quantifies the number of vacant housing units within neighborhoods, has an average of 76,030. The substantial standard deviation of 19,589 suggests disparities in housing vacancy rates across different areas, potentially indicative of varied economic conditions. These vacancies could offer a dual narrative: on one hand, they may represent a surplus in housing that could be leveraged by the short-term rental market; on the other, they could signal underlying economic challenges that merit further investigation by urban planners and policymakers. Available housing units could also indicate supply levels; more vacant units might reduce prices due to increased competition among sellers, whereas fewer vacancies could lead to higher prices.

'Median Household Income' can serve as a socioeconomic barometer of conditions within a borough. The average income is 78,444 and has relatively low variance with a standard deviation of 13,852. Median neighborhood income can be seen as a proxy for local economic health; areas with higher incomes may have more expensive properties and thus command a higher Airbnb price, while the opposite may be true for areas with low median income. We can speculate that median neighborhood income may also capture other neighborhood characteristics related to income such as cleanliness, crime rates, and other socio-economic indicators.

These statistics offer a snapshot of the economic environment in which Airbnb operates, suggesting that proximity to tourism centers, the prevalence of vacation housing, and the financial well-being of households are influential factors. High tourism proximity likely correlates with increased demand for short-term rentals, whereas the density of vacation housing units could indicate market maturity and competitiveness. Concurrently, median household income levels might influence the pricing strategies of listings and determine the economic impact of short-term rentals on local communities.


Figure 1: Review per Month against Price.

We can curiously observe a negative relationship by reviews per month and price, suggesting that listings with a higher amount of reviews per month charge less. This could indicate a possible long run strategy in which hosts set relatively low prices in order to build reputation, but it can also signal a bidirectional relationship in customers are more likely to leave a review for listings with a lower price.


Figure 2: Box plot of calculated host listings against price.

Figure 2 suggests a non linear relationship between the amount of listings a host owns and the prices they charge. Hosts with a few amount of listings will charge lower and lower prices until hosts with about 10+ listings diverge and charge far higher prices with higher variance. The likely explanation for this is that hosts who own more listings also maintain nicer properties, have better reputation, better services, and/or better location and are thus able to get away with charging a premium. Regardless less of the specific explanation what we can take away from the plot is that host characteristic do have some significant impact on price.

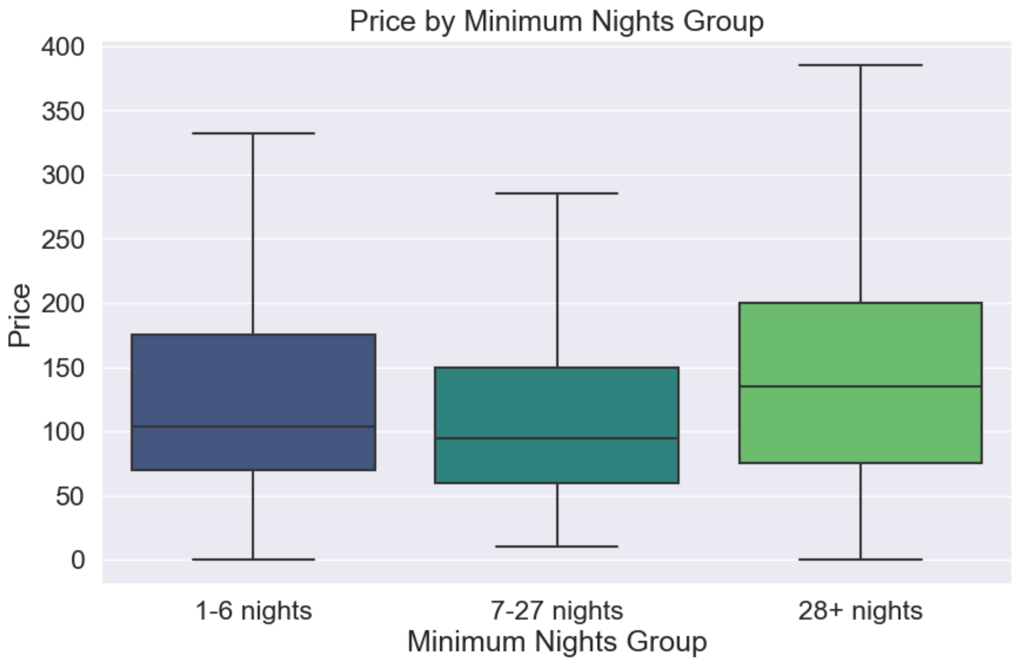


Figure 3: Box plot of minimum night groups against price.

Figure 3 suggests that minimum nights likely has some impact on price. The results can be explained economically quite well. A higher price for short stays (1-6 nights) can be explained by the fact that listings with short minimum stays may attract guests looking for brief stays, possibly willing to pay a premium for flexibility. Hosts might set higher prices to maximize revenue per night to offset the higher turnover costs, such as cleaning and administration, that come with short-term rentals. Mid length Stays (7-27 nights) might be priced slightly lower as they attract a different segment of the market—guests seeking accommodations for longer durations like extended vacations or business trips. The slightly lower prices could be a strategy to make these properties more attractive and ensure occupancy, given that guests staying longer are less likely to pay a premium for each night. Additionally, the reduced turnover for such properties can lead to lower operational costs per booking, allowing hosts to offer more competitive rates. Lastly long stays (28+ nights) could be priced higher due to their appeal to guests looking for long-term accommodations, but who are not able to secure tenancy in New York.

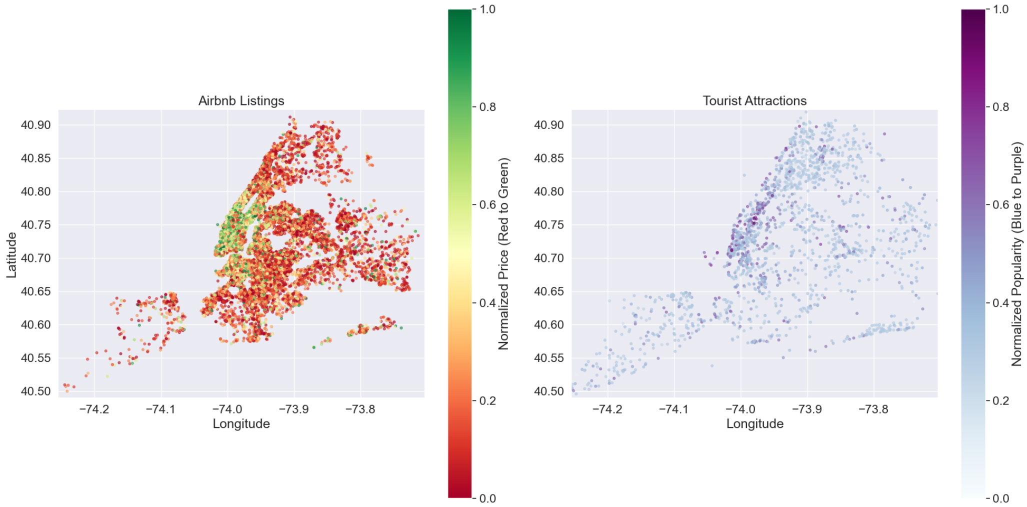


Figure 4: Map of airbnb listings with normalized price (left) and map of tourist attractions with normalized popularity (right).

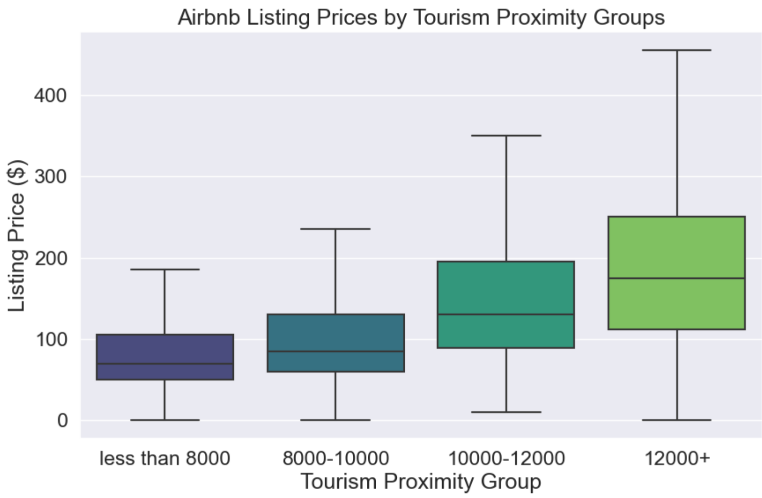


Figure 5: Grouped tourism proximity against price.

Observing figure 4 and 5 it appears that price is strongly correlated with tourist attractions. We can see that popular tourist destinations pretty much everywhere in NYC are almost always surrounded by higher priced listings. Locations with more tourist attractions also appear to have more Airbnbs, while locations that are sparse of tourist destinations also have blank spots in listings. Economically, this makes sense, as we would expect that in areas with more attractions there would be a higher economic incentive to host an Airbnb.

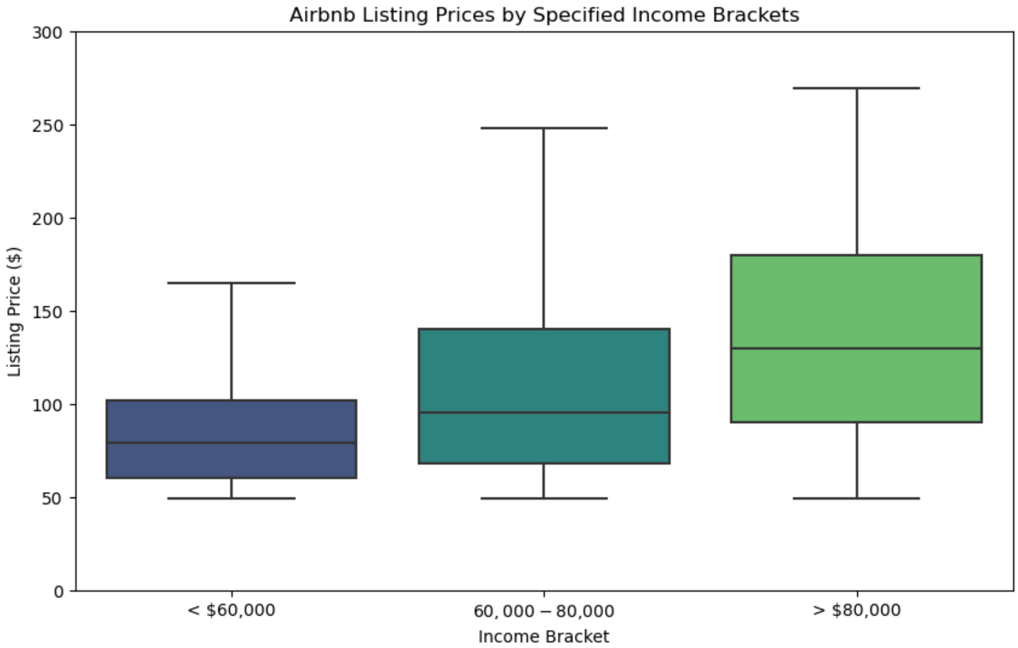


Figure 6: Map of airbnb listings with normalized price (left) and map of tourist attractions with normalized popularity (right).

Figure 6 also depicts a linear relationship between the price of Airbnb listings and median household income of a neighborhood. This potentially encapsulates underlying socioecnomic characteristics that might be attractive to Airbnb's customerbase, as speculated upon earlier.

# 4 Results

## 4.1 OLS Regressions

| | *Dependent variable:* log_price |
|---|---|
| | (1) |
| Normalized_Tourism_Proximity | 2.837*** |
| | (0.047) |
| Normalized_VacHUs | 0.280*** |
| | (0.014) |
| Normalized_availability_365 | 0.340*** |
| | (0.008) |
| Normalized_host_listings_count | 8.215*** |
| | (0.645) |
| Normalized_host_listings_count_squared | -7.625*** |
| | (0.660) |
| Normalized_minimum_nights | -0.370*** |
| | (0.020) |
| Normalized_reviews_per_month | -1.078*** |
| | (0.087) |
| const | 3.085*** |
| | (0.019) |
| entire_homeapt | 1.198*** |
| | (0.017) |
| listings_count_squared_x_availability_365 | 0.000*** |
| | (0.000) |
| listings_count_x_availability_365 | -9.484*** |
| | (0.875) |
| private_room | 0.440*** |
| | (0.017) |
| Observations | 35973 |
| $R^2$ | 0.539 |
| Adjusted $R^2$ | 0.538 |
| Residual Std. Error | 0.459 (df=35961) |
| F Statistic | 3815.568*** (df=11; 35961) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

We will only show the results of our final OLS regression here for brevity but will proceed to explain the remaining regressions. The other regressions are included in the appendix.

In the first set of OLS regressions analyzing neighborhood characteristics affecting Airbnb prices in New York City, the increment in $R^2$ by only 0.03 upon adding neighborhood dummy variables suggested limited additional explanatory power. However, high coefficients for tourism proximity across models and the subsequent VIF scores (1.99 for tourism proximity and extraordinarily high values exceeding 1170 for Vacant Housing Units and infinite for median household income and borough dummies) indicated significant multicollinearity, affirming tourism proximity as a dominant explanatory factor. This was evidenced by a higher $R^2$ in models including tourism proximity alone (0.196) compared to those incorporating broader neighborhood variables (0.116), highlighting its substantial impact over other location-based attributes. Moreover, the negative coefficient for tourism proximity squared suggested diminishing returns on price increases beyond a normalized score of about 0.6. These results, backed by solid p-values and F-statistics, underline the significant influence of tourism proximity on Airbnb pricing, providing crucial insights in the initial regression analysis set.

In the second set of OLS regressions examining listing characteristics, we confirmed that entire homes/apartments command significantly higher prices than private rooms, which in turn command a premium over shared rooms, aligning with prior visual analyses. Notably, there is a persistent negative relationship between the number of minimum nights required and listing prices, suggesting that hosts may lower prices to attract longer-term stays, thereby reducing operational costs such as cleaning and guest communication. Interestingly, the positive coefficient for the squared term of minimum nights indicates a price premium at higher thresholds, likely aimed at longer-term tourists unable to secure traditional accommodations. Reviews per month and the total number of reviews negatively impact price, although interaction terms between reviews per month and non-shared room types show unusually high positive coefficients, potentially indicating a premium on popular entire homes and private rooms. Minimal changes in $R^2$ from these interactions led to their exclusion in later models to clarify underlying effects. Additionally, the effect of tourism proximity varies by room type, being more pronounced for private rooms, possibly due to greater flexibility for hosts in tourist-dense areas to rent out parts of their residences. The regression models maintain statistical robustness, evidenced by significant standard errors, t-stats, p-values, and F-stats, validating the findings' reliability in depicting how listing specifics drive pricing dynamics in New York City's Airbnb market.

The our third sets of regressions focusing on host characteristics revealed nuanced effects on Airbnb pricing. Initially, the host listings count shows a general positive correlation with price, prompting the inclusion of a squared term in subsequent regressions to capture the relationship's complexity. The results confirm that hosts with fewer listings tend to charge less, while those with more listings, especially when squared, exhibit the ability to command higher prices, suggesting scale advantages. This is further complicated by the interaction between host listings count and monthly reviews, where a negative coefficient indicates diminishing returns on price with increasing listings and reviews. However, a positive coefficient for the squared interaction term suggests that beyond a certain threshold, hosts can leverage their extensive portfolio and high review volumes to demand premium pricing. Despite these statistically significant findings, the $R^2$ value of 0.035 in the expanded regression model indicates that host characteristics and their interactions explain only a small fraction of the variance in log price, highlighting their limited economic impact compared to other factors like room characteristics. This modest influence is somewhat unexpected, contrasting with prior research that suggests host attributes are as significant as room features, potentially due to our dataset's less comprehensive host-related information and the complexities introduced by interaction terms.

Lastly, using the variables and results we had uncovered in our previous regression and visual analysis we constructed our final regression as shown above. Our final regression and specification is outputted below. All the variables for our final regression are statistically and economically significant, have t-stats over 10 and statistically significant p-values. We can note that we have an $R^2$ of 0.53, indicating that we have explained 53.8% of pricing variance within our test group. This result is quite good considering that from an economic standpoint there will likely be quite a bit of irrationality in airbnb pricing decisions since most host's are amateurs. This $R^2$ is in line with other comprehensive airbnb research into pricing as well such as the Barcelona study (Toader et al, 2020) who had $R^2$ of 0.488 and the Canadian airbnb study which explained between 48.8% to 68.8% of the variance in various cities (Gibbs et al, 2018). This is our final OLS preferred specification. We can also note the adjusted $R^2$ to be about the same, providing some indication that our variable choice was appropriate. The mean squared error (MSE) is approximately 0.21 and root mean squared error is about 0.447. Considering that log price has an average of 4.710729 and standard deviation of about 0.7, this MSE is decent and about in line with our $R^2$. The mse suggest the model's predictions are about 46% off from the actual price when considering smaller errors. Again this is somewhat large but not unexpected considering the variance in price and our variables.

## 4.2 Decision Trees

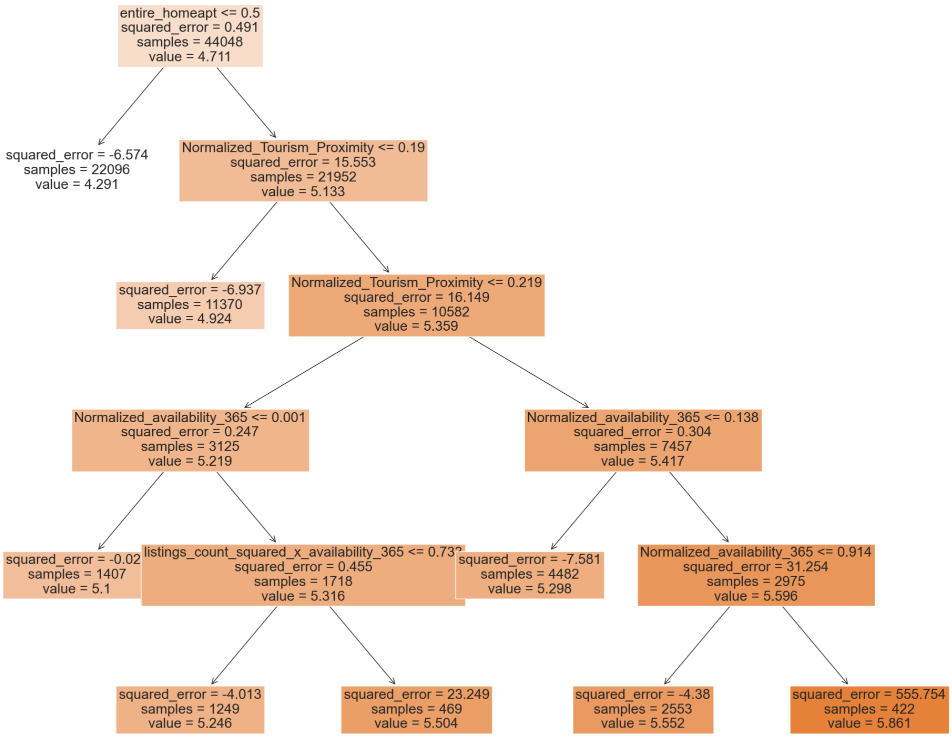


Figure 7: Decision tree using variables from preferred specification.

The decision tree analysis, as illustrated in Figure 7, reinforces our findings that room type and tourism proximity are the primary determinants of Airbnb listing prices. It demonstrates an initial splitting based on whether a listing is an entire home or apartment, which is economically intuitive since such listings typically offer more privacy and space—a valued commodity in the hospitality industry. Subsequently, the tree prioritizes normalized tourism proximity, reflecting the economic principle that locations closer to popular tourist destinations can command higher prices due to increased demand.

Interestingly, the decision tree prunes other variables early, indicating their relative non-importance compared to the primary splitting variables. This is in line with our prior findings where room type and tourism proximity emerged as substantial determinants of pricing. The mean squared error of the decision tree stands at approximately 0.29, which, while higher than that of our linear regression models, suggests room for improvement in capturing the complexity of pricing influences.

Furthermore, the tree's decision to split on availability 365 highlights the existence of a non-linear relationship between availability and price that was not evident from our linear models.

This suggests that although availability may not have a straightforward linear correlation with price, it plays a significant role in pricing strategy when considered through the lens of market dynamics and consumer behavior, particularly in the context of short-term rental availability and its perceived scarcity or abundance.

This analysis provides valuable insights into the economic interactions between listing characteristics and their influence on price, affirming the utility of decision tree models in uncovering complex patterns within the data that linear approaches may overlook.

## 4.3 Random Forest Model and Importance Matrix

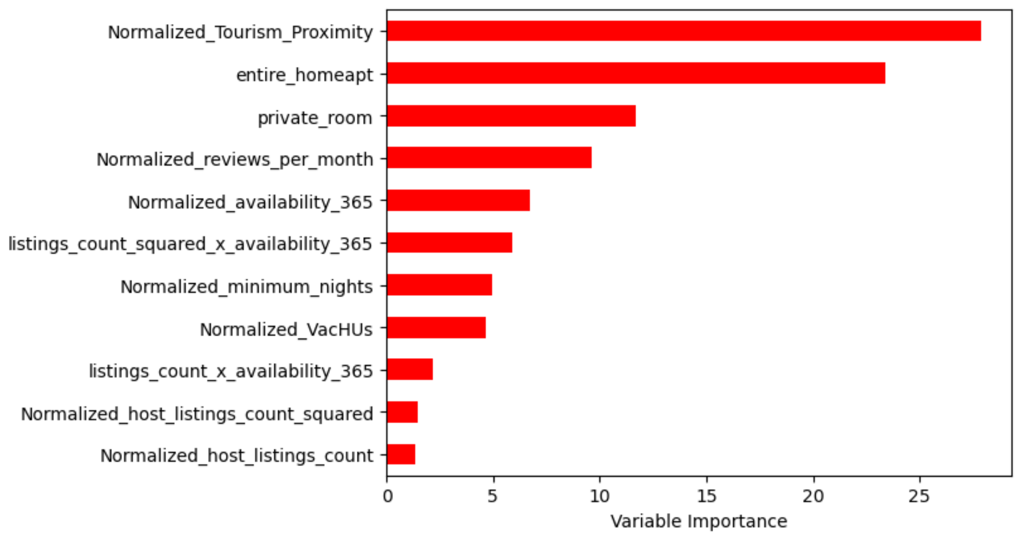


Figure 8: Importance matrix generated from random forest model.

The Random Forest model's variable importance plot offers a perspective that mostly aligns with our previous analysis on the determinants of Airbnb pricing. Normalized tourism proximity accounts for roughly 27% of the model's predictive power, representing the impact of location desirability and access to tourist attractions. Close behind, with an importance of approximately 23%, is our dummy variable for entire homes and apartments, then private room's. This confirms the importance of room type as a listing characteristic and price determinant.

The importance plot also reveals that reviews per month exert a stronger influence on pricing than anticipated, suggesting that a listing's popularity and guest turnover rate may be more closely related to price setting than initially posited. This observation aligns with findings from the comprehensive regression and may reflect a market where higher review volumes correlate with increased consumer trust and, consequently, a willingness to pay more.

With a mean squared error of around 0.03, the Random Forest model demonstrates superior predictive accuracy relative to the earlier regression models. Considering the variance observed in log price (mean of 4.710729 and standard deviation of 0.700714), such a low mean squared error is indicative of the selected variables' substantial explanatory power. It economically validates the robustness of the factors chosen for the analysis, suggesting that these variables collectively capture a significant portion of the price variability in the Airbnb marketplace. The precision of the Random Forest model, alongside the corroborating results from the importance matrix, provides strong empirical evidence for the influence of each variable on the pricing dynamics within New York City's Airbnb listings.

# 5 Conclusion

The results of our study have confirmed several findings from previous research as well as uncovered some novel findings of our own. We have confirmed previous research that listings characteristics such as room type, and minimum nights have significant effects on price, notably we have confirmed the unusual finding from some research that a greater amount of active reviews for a listing is correlated with a lower listing price.

We have also found that the number of listings a host owns- a likely proxy for experience, also have a significant positive effect on price through interaction terms, and that neighborhood characteristics such as median income and vacant housing units have a positive effect on price as well. Interestingly, we noted that vacant housing units provided a bigger explanation for price than median neighborhood income, suggesting that the forces of supply and demand are at play for Airbnb's in New York City. While previous research examining location based factors on price has limited their spatial analysis to using straight line distance's to a handful of notable attraction such as beaches, city halls, and city centers. Our approach to using tourist attractions data and their associated reviews to calculate popularity and tourism proximity scores for each listing is novel and far more comprehensive and gave us our most significant result; that a listing's proximity to a larger number of popular tourist destinations is the biggest determinant of price. Additionally, our method of collecting tourist attractions could be improved upon as suggested in our data section by expanding our search terms or securing alternative forms of tourism data. We were also forced to drop listings with minimum nights over 28 days due to our limited data on them over our dummy variables, which while only dropping 4811 listings was undesirable. Future studies may wish to make use of more data or find other ways to combat numerical errors introduced by the limited sample size.

Nevertheless all our findings were statistically significant and our OLS regressions explained about 53.8% of variance with relatively little error, which is in line with similar research into Airbnb pricing determinants (Toader et al., 2020; Gibbs et al,. 2017; Deboosere et al., 2019).

The story that emerges from our findings is that the primary users of airbnb in New York City are short term tourists with low elasticities of demand and that a high portion of airbnbs are managed by economical hosts. The consistent findings that reviews per month has a negative correlation with price suggests a bidirectional relationship and/or more economical hosts seeking to take advantage of long run strategies to build reputation in an incipient market. Though our findings only weakly supported this theory, a time series causal analysis would uncover the exact reasoning behind it. Furthermore, an additional limitation of the study was lacking detailed host characteristics which could've allowed us to potentially identify economical hosts better and provided additional insight into our data. Establishing causal relationships between our different variables and price would have furthered our understanding of pricing determinants as well.

In conclusion our study was comprehensive and provided important insight to support existing research, while our findings for tourism attractions effect on price expanded the understanding of Airbnb pricing determinants. A future study may seek to determine if the effects of tourism attractions and their popularity are generalize across different municipalities.

# References

[1] Voltes-Dorta, A., & Sánchez-Medina, A. (2020, September 11). Drivers of airbnb prices according to property/room type, season and location: A regression approach. *Journal of Hospitality and Tourism Management*. Retrieved from https://www.sciencedirect.com/science/article/pii/S1447677020302023

[2] Zhang, Z., Chen, R. J. C., Han, L. D., & Yang, L. (2017, September 14). Key factors affecting the price of Airbnb listings: A geographically weighted approach. *MDPI*. Retrieved from https://www.mdpi.com/2071-1050/9/9/1635

[3] Perez-Sanchez, V. R., Serrano-Estrada, L., Marti, P., & Mora-Garcia, R.-T. (2018, December 5). The what, where, and why of airbnb price determinants. *MDPI*. Retrieved from https://www.mdpi.com/2071-1050/10/12/4596

[4] Gutiérrez, J., García-Palomares, J. C., Romanillos, G., & Salas-Olmedo, M. H. (2017, May 12). The eruption of airbnb in tourist cities: Comparing spatial patterns of hotels and peer-to-peer accommodation in Barcelona. *Tourism Management*. Retrieved from https://www.sciencedirect.com/science/article/pii/S0261517717301036

[5] 2020 census. 2020 Census -DCP. (n.d.). Retrieved from https://www.nyc.gov/site/planning/planning-level/nyc-population/2020-census.page

[6] United States Census Bureau. (2019). Retrieved from https://data.census.gov/

[7] Teubner, T., Hawlitschek, F., & Dann, D. (2017). Price determinants on Airbnb: How reputation pays off in the sharing economy. *Journal of Self-Governance & Management Economics*, 5(4), 53–80.

[8] Wang, D., & Nicolau, J. L. (2017). Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb.com. *International Journal of Hospitality Management*, 62, 120–131.

[9] Buchanan, L., Katz, J., Washington, E., & Taylor, R. (2023, October 30). An extremely detailed map of New York City neighborhoods. *The New York Times*. Retrieved from https://www.nytimes.com/interactive/2023/upshot/extremely-detailed-nyc-neighborhood-map.html

[10] Valentin Toader, Adina Letiţia Negruşa, Oana Ruxandra Bode & Rozalia Veronica Rus (2022) Analysis of price determinants in the case of Airbnb listings, *Economic Research-Ekonomska Istraživanja*, 35:1, 2493-2509, DOI: 10.1080/1331677X.2021.1962380

[11] Kakar, V., Voelz, J., Wu, J., & Woo, J. (2017, August 25). The visible host: Does race guide airbnb rental rates in San Francisco?. *Journal of Housing Economics*. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1051137716301309`

[12] Gibbs, C., Guttentag, D., Gretzel, U., Morton, J., & Goodwill, A. (2018). Pricing in the sharing economy: a hedonic pricing model applied to Airbnb listings. *Journal of Travel & Tourism Marketing*, 35(1), 46–56. DOI: `10.1080/10548408.2017.1308292`

[13] Deboosere, R., Kerrigan, D. J., Wachsmuth, D., & El-Geneidy, A. (2019). Location, location and professionalization: a multilevel hedonic analysis of Airbnb listing prices and revenue. *Regional Studies, Regional Science*, 6(1), 143–156. DOI: `10.1080/21681376.2019.1592699`

Regression set 1 (neighborhood variables)

| | *Dependent variable: $log_price$* | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Normalized$_{M}edian_{I}ncome$ | 0.154$^{***}$ | 0.188$^{***}$ | |
| | (0.021) | (0.020) | |
| Normalized$_{T}ourism_{P}roximity$ | 6.018$^{***}$ | 6.293$^{***}$ | |
| | (0.190) | (0.192) | |
| Normalized$_{T}ourism_{P}roximity_{S}quared$ | -4.965$^{***}$ | -5.228$^{***}$ | |
| | (0.409) | (0.409) | |
| Normalized$_{V}acHUs$ | 0.028 | 0.356$^{***}$ | |
| | (0.027) | (0.011) | |
| borough$_{B}rooklyn$ | | -0.005 | 0.327$^{***}$ |
| | | (0.011) | (0.021) |
| borough$_{M}anhattan$ | | -0.177$^{***}$ | 0.735$^{***}$ |
| | | (0.007) | (0.021) |
| borough$_{Q}ueens$ | | 0.028$^{**}$ | 0.151$^{***}$ |
| | | (0.013) | (0.022) |
| borough$_{S}tatenIsland$ | | 0.379$^{***}$ | 0.133$^{***}$ |
| | | (0.022) | (0.040) |
| const | 3.657$^{***}$ | 3.412$^{***}$ | 4.242$^{***}$ |
| | (0.018) | (0.027) | (0.020) |
| Observations | 44048 | 44048 | 44048 |
| $R^2$ | 0.196 | 0.199 | 0.116 |
| Adjusted $R^2$ | 0.196 | 0.199 | 0.116 |
| Residual Std. Error | 0.628 (df=44043) | 0.627 (df=44041) | 0.659 (df=44043) |
| F Statistic | 2691.507$^{***}$ (df=4; 44043) | 1822.630$^{***}$ (df=6; 44041) | 1439.035$^{***}$ (df=4; 44043) |
| *Note:* | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Regression set 2 (listing variables)

| | *Dependent variable: $log_{p}rice$* | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Normalized$_{M}edian_{I}ncome$ | 0.154*** | 0.188*** | |
| | (0.021) | (0.020) | |
| Normalized$_{T}ourism_{P}roximity$ | 6.018*** | 6.293*** | |
| | (0.190) | (0.192) | |
| Normalized$_{T}ourism_{P}roximity_{S}quared$ | -4.965*** | -5.228*** | |
| | (0.409) | (0.409) | |
| Normalized$_{V}acHUs$ | 0.028 | 0.356*** | |
| | (0.027) | (0.011) | |
| borough$_{B}rooklyn$ | | -0.005 | 0.327*** |
| | | (0.011) | (0.021) |
| borough$_{M}anhattan$ | | -0.177*** | 0.735*** |
| | | (0.007) | (0.021) |
| borough$_{Q}ueens$ | | 0.028** | 0.151*** |
| | | (0.013) | (0.022) |
| borough$_{S}tatenIsland$ | | 0.379*** | 0.133*** |
| | | (0.022) | (0.040) |
| const | 3.657*** | 3.412*** | 4.242*** |
| | (0.018) | (0.027) | (0.020) |
| Observations | 44048 | 44048 | 44048 |
| $R^2$ | 0.196 | 0.199 | 0.116 |
| Adjusted $R^2$ | 0.196 | 0.199 | 0.116 |
| Residual Std. Error | 0.628 (df=44043) | 0.627 (df=44041) | 0.659 (df=44043) |
| F Statistic | 2691.507*** (df=4; 44043) | 1822.630*** (df=6; 44041) | 1439.035*** (df=4; 44043) |

*Note:* *p<0.1; **p<0.05; ***p<0.01

| | *Dependent variable: $log_price$* | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Normalized$_{a}vailability_{3}65$ | 0.149*** | 0.184*** | 0.178*** |
| | (0.011) | (0.011) | (0.011) |
| Normalized$_{h}ost_{l}istings_{c}ount$ | 0.766*** | -5.262*** | -4.839*** |
| | (0.049) | (0.399) | (1.008) |
| Normalized$_{h}ost_{l}istings_{c}ount_{s}quared$ | | 6.093*** | 5.485*** |
| | | (0.400) | (1.021) |
| Normalized$_{m}inimum_{n}ights$ | 0.825*** | 0.779*** | 0.770*** |
| | (0.066) | (0.066) | (0.066) |
| Normalized$_{m}inimum_{n}ights_{s}quared$ | -1.267*** | -1.187*** | -1.282*** |
| | (0.088) | (0.088) | (0.088) |
| Normalized$_{r}eviews_{p}er_{m}onth$ | -1.062*** | -1.063*** | -0.157 |
| | (0.128) | (0.128) | (0.145) |
| const | 4.627*** | 4.636*** | 4.628*** |
| | (0.007) | (0.007) | (0.007) |
| listings$_{c}ount_{s}quared_{xa}vailability_{3}65$ | | | -0.000** |
| | | | (0.000) |
| listings$_{c}ount_{s}quared_{xm}onthly_{r}eviews$ | | | 202.946*** |
| | | | (16.208) |
| listings$_{c}ount_{xa}vailability_{3}65$ | | | 2.268* |
| | | | (1.265) |
| listings$_{c}ount_{xm}in_{n}ights$ | | | 12.282*** |
| | | | (1.415) |
| listings$_{c}ount_{xm}onthly_{r}eviews$ | | | -199.405*** |
| | | | (15.929) |
| Observations | 35973 | 35973 | 35973 |
| $R^2$ | 0.020 | 0.026 | 0.035 |
| Adjusted $R^2$ | 0.020 | 0.026 | 0.034 |
| Residual Std. Error | 0.668 (df=35967) | 0.666 (df=35966) | 0.663 (df=35961) |
| F Statistic | 147.089*** (df=5; 35967) | 161.943*** (df=6; 35966) | 117.157*** (df=11; 35961) |
| *Note:* | | | *p<0.1; **p<0.05; ***p<0.01 |