Decoding Airbnb: Price Influences in New York City

Jessica Wang

Introduction

In this project, we undertake an exploratory analysis of the Airbnb Open Data available on Kaggle (https://www.kaggle.com/datasets/arianazmoudeh/airbnbopendata/data). Airbnb is an online marketplace for lodging, homestays and vacation rentals. It provides a platform for hosts to offer a variety of accommodations to travelers. This dataset offers a rich compilation of Airbnb listings and reviews data in New York City from 2008 to 2022, covering various factors such as listing price, location, ratings, and other relevant attributes.

The main question driving our research is: "What factors significantly influence the price of an Airbnb house listing?" We hypothesize that a variety of factors, from the geographical location of a listing to the ratings left by previous customers, have a significant impact on how hosts set their prices. Our goal is to analyze the dataset to identify any patterns, correlations, and potentially causal relationships between these factors and the prices of Airbnb listings.

The significance of this project lies in its potential to provide insights that could be useful for Airbnb hosts looking to set competitive prices, for guests trying to find the best value for their stays, and for policymakers interested in understanding the impact of the sharing economy on traditional lodging and tourism industries. By examining the diverse factors that can affect Airbnb pricing, this project aims to contribute valuable information to the ongoing discussion about the economics of sharing economy platforms like Airbnb.

Methods

This dataset comprises 102,599 observations across 26 variables. An initial assessment reveals several issues. Most notably, numerous variables exhibit missing data due to the unprocessed nature of the information sourced from the Airbnb open database. By standardizing these missing values to the NA format, we can determine the missing data rate for each variable. The results show that the missing data rates for all variables are relatively low; hence, we remove the observations with missing data.

Moreover, several variables present challenging data types for manipulation. For example, housing prices are recorded as character strings with a prefixed dollar sign. We address this by stripping the dollar sign and converting the prices to numerical values. Likewise, we transform categorical variables, such as the listing neighborhood, from strings to factors where suitable.

A detailed examination of each variable using bar plots and histograms highlights further anomalies. The "minimum nights" variable, indicating the required minimum stay, included implausible figures, such as negative numbers or excessively high values up to 5,645 nights. We eliminate negative values and any figures exceeding the third quartile (75%) plus 1.5 times the IQR. Additionally, we examine the neighborhood groups and notice that one of the groups is a typo. Specifically, both "brookln" and "Brooklyn" represent the same neighborhood group. Thus, we update all "brookln" neighborhoods to "Brooklyn". The last variable we made modified is "available 365". This variable represents the number of days this listing is available for in the coming year. As such, any negative values or values above 365 days are impossible and thus removed.

Through the barplots and histograms, we also uncover some additional information that does not require modification but can be helpful in future analysis. We observe that the number of reviews for the listings is heavily right-skewed with the majority of listings having few reviews but also several listings having as high as 1,024 number of total reviews. These outliers are retained since they are likely to represent natural variations in the total number of reviews instead of data collection errors. Furthermore, the distributions of

listing prices and listing rates (on a Likert scale of 1 to 5) appear fairly uniform and thus do not require additional processing.

Characteristic	**N = 68,428**[1]
Host Verified	
unconfirmed	34,147 (50%)
verified	34,281 (50%)
Minimum Nights Requirement	
1	20,116 (29%)
2	20,316 (30%)
3	13,459 (20%)
4	5,408 (7.9%)
5	4,753 (6.9%)
6	1,216 (1.8%)
7	2,846 (4.2%)
8	195 (0.3%)
9	119 (0.2%)
Boroughs	
Bronx	2,053 (3.0%)
Brooklyn	29,161 (43%)
Manhattan	26,943 (39%)
Queens	9,528 (14%)
Staten Island	743 (1.1%)
Instant Bookable	
FALSE	34,318 (50%)
TRUE	34,110 (50%)

[1] n (%)

Characteristic	**N = 68,428**[1]
Ratings	
1	6,045 (8.8%)
2	15,474 (23%)
3	15,661 (23%)
4	15,660 (23%)
5	15,588 (23%)
Room Type	
Entire home/apt	34,514 (50%)
Hotel room	97 (0.1%)
Private room	32,397 (47%)
Shared room	1,420 (2.1%)
Cancellation Policy	
flexible	22,805 (33%)
moderate	22,845 (33%)
strict	22,778 (33%)
Construction Year	2,012 (2,008, 2,018)
Listing Price	626 (341, 914)
Service Fee	125 (68, 183)
Total Number of Reviews	13 (4, 42)
Number of Days This Listing is Available	87 (3, 240)
Reviews per Month	0.98 (0.28, 2.33)

[1] n (%); Median (IQR)

The processed key variables are summarized in the table above. Discrete variables are analyzed for their frequency distribution, with percentages. Continuous variables are summarized using their median values and interquartile ranges in the format: Median (Lower Quartile, Upper Quartile).

Results

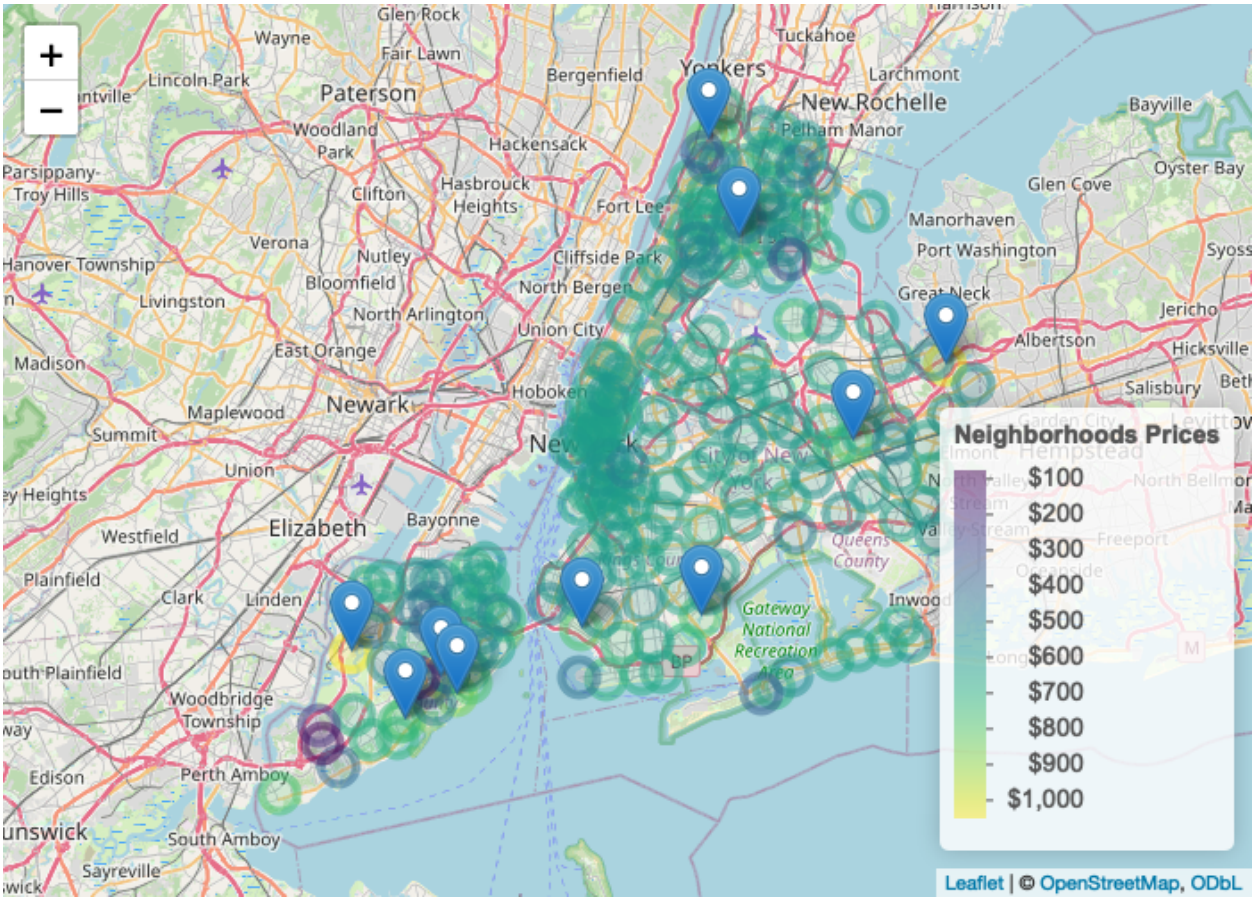
In this section, we proceed to study several key variables and their relationships with the Airbnb housing price.

Table 3: Top 10 Neighbourhoods by Average Price

Neighborhood	Average Listing Price
New Dorp	1048.00
Chelsea, Staten Island	1042.00
Little Neck	1010.00
New Dorp Beach	856.71
Riverdale	822.42
Bay Terrace, Staten Island	800.00
East Morrisania	799.78
Jamaica Hills	796.00
Mill Basin	775.14
Bath Beach	772.66

The figure above displays boxplots of Airbnb listing prices grouped by the five boroughs of New York City. The plot suggests that the median listing price does not differ substantially among the boroughs. Moreover, the price ranges are also comparable across the boroughs. To gain further insights into the relationship between geographic location and listing price, we refine our analysis by increasing the granularity, moving from boroughs to individual neighborhoods within these boroughs.

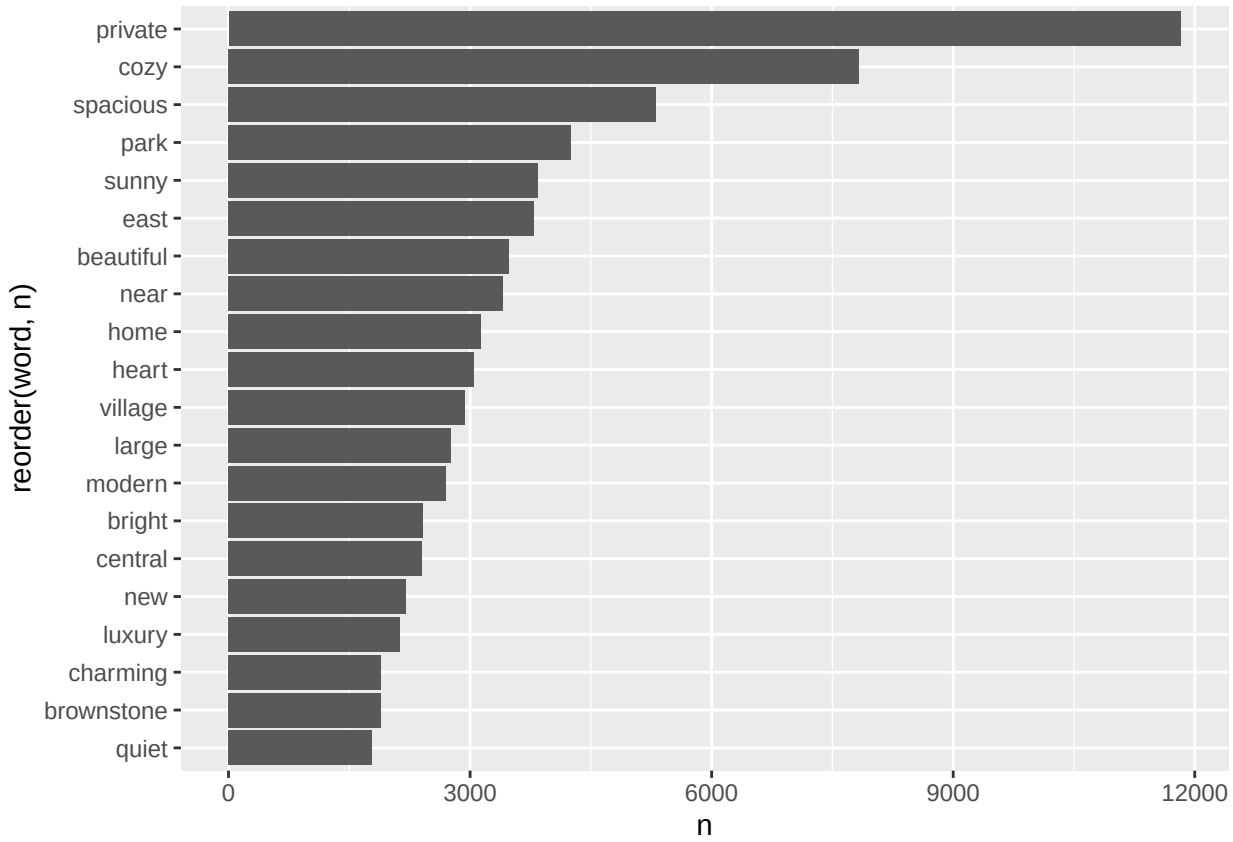
The table above identifies the ten neighborhoods with the highest average listing price. In order to better visualize this information geographically, a map by average listing price is displayed below.



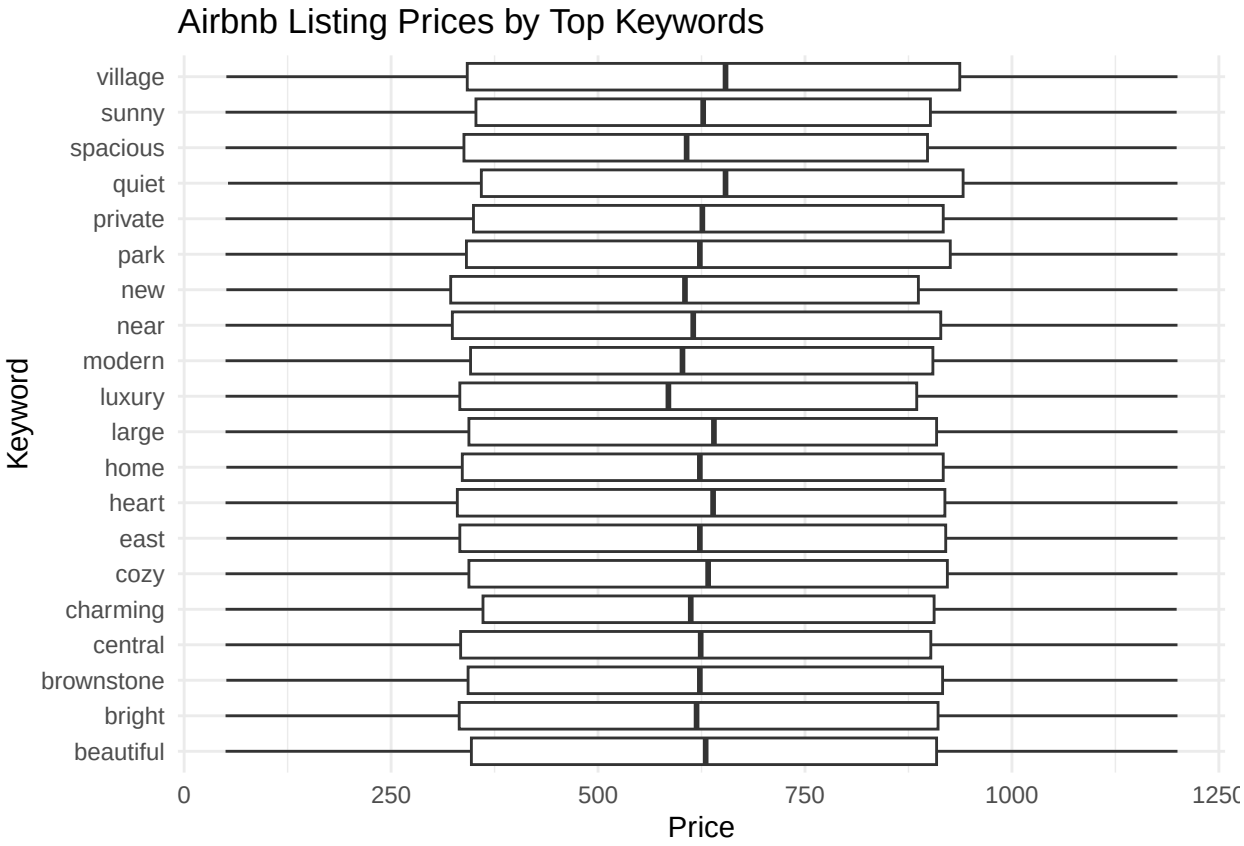
The leaflet map presented above illustrates the average prices of Airbnb listings by neighborhood in New York City. An interactive version of this map can be found on the visualization tab of the project website. Circle markers are color-coded according to the average price. Darker colors (such as purple) correspond to lower average prices, while brighter colors (like yellow) signify higher average prices. Additionally, ten blue pins highlight the neighborhoods with the top ten highest average listing prices. The map reveals a higher density of Airbnb listings in areas closer to the city center or downtown, such as Lower Manhattan, as opposed to more peripheral areas like Staten Island. This observation aligns with the typical demand patterns in popular tourist destinations. However, the top ten neighborhoods with highest average prices are situated in somewhat suburb areas. This pattern could be due to the fact that these locations have fewer competitions and offer unique accommodations that cater to niche markets or provide luxury experiences. Consequently, while the dense urban core of the city offers a high number of listings likely aiming at a broad market, the more residential and less saturated suburban areas provide distinctive options with potentially higher prices.

The boxplot above visualizes the distribution of the Airbnb listing prices per night across different minimum stay requirements, ranging from 1 to 9 nights. Observing the median prices, there is no clear upward or downward trend associated with the length of the minimum stay requirement. The medians appear relatively stable across different categories. However, it is noteworthy that the median listing price for accommodations requiring a 9-night minimum stay is comparatively lower than for other categories. This decrease suggests that hosts may offer more competitive prices for longer stays to attract guests willing to commit to this duration.

The facet plot displays the relationship between the total number of reviews and Airbnb listing prices across five neighborhoods with the highest count of listings. The visualization indicates that for neighborhoods like Hell's Kitchen and Williamsburg, there is a slight positive trend, indicating that listings with more reviews tend to have higher prices. In contrast, neighborhoods like Bedford-Stuyvesant and Bushwick show a flatter regression line, suggesting a weaker relationship between the number of reviews and price. Harmlem has a slightly negative trend, implying that more reviews could correlate with lower prices.There is no single clear pattern dominating across the neighborhoods, which suggests that while there may be some local trends within each neighborhood regarding reviews and pricing, they do not appear to be strong or consistent across different areas.



The bar chart above displays the top 20 most frequently used words in the names of Airbnb listings in New York City, after excluding common English stopwords and specific terms that are overly generic or location-based, such as 'NYC', 'room', 'bedroom', 'bed', 'Manhattan', 'Williamsburg', and 'Brooklyn'. Additionally, any numeric terms have been filtered out to focus solely on descriptive words. The words are presented in descending order of frequency, showcasing that descriptors such as "private", "cozy", and "spacious" are most popular among hosts when naming their listings.



In our investigation into how specific keywords in Airbnb listing titles affect price distribution, we observed some counterintuitive trends regarding the perceived value and pricing strategies among hosts. Notably, from the figure above we see that listings featuring the keyword 'luxury' surprisingly is not associated with highest prices; instead, they exhibit lower median prices compared to other popular keywords. This could be attributed to several factors:

1. Overuse of the Term 'Luxury': The term 'luxury' may be overused in marketing contexts, often being applied to a wide range of properties, including those that do not necessarily offer true luxury experiences. This overuse can dilute the term's effectiveness in signaling genuine high-end offerings.

2. Consumer Expectations: Travelers might be skeptical of listings labeled 'luxury' without sufficient validation through reviews or ratings, especially if the price seems incongruous with their expectations for luxury accommodations. This skepticism might drive down the median price as potential guests opt for listings that offer clearer value or whose luxury claims are substantiated by strong reviews.

On the other hand, listings that incorporate 'village' or 'quiet' in their titles command higher median prices. These keywords seem to attract travelers for several reasons:

1. Desire for Specific Experiences: Keywords like 'village' and 'quiet' may appeal strongly to those seeking a particular type of experience – one that promises a peaceful setting. This appeal can drive up demand and, consequently, prices, as these terms align well with the desires of guests looking for retreats away from bustling city centers.
2. Limited Availability and High Demand: Listings that can credibly claim to offer 'quiet' environments in urban settings such as New York City are likely scarce but in high demand. Therefore, this limited availability can make these listings more valuable.

These findings underscore the importance of keyword choice in listing titles as a significant factor in pricing strategy. They highlight how certain terms can either enhance or detract from perceived value, influenced by overuse, consumer expectations, or the genuine appeal of the experience promised by the keyword.

Table 4: Linear Model Summary

	Estimate	StdError	tValue	Pr
(Intercept)	636.901	4.222	150.846	0.000
Total Number of Reviews	0.019	0.024	0.783	0.434
Minimum Nights Requirement	-1.624	0.792	-2.052	0.040
Customer Ratings	-2.105	0.990	-2.126	0.033

Given the findings identified above regarding the influence of the minimum nights requirement and the total number of reviews on the listing price, we fitted a linear model with the listing price as the response variable, using the minimum nights requirement and the total number of reviews as predictor variables. Additionally, we included customer ratings as another predictor, since they can significantly reflect guests' satisfaction and may potentially impact a listing's pricing strategy.

The model's intercept is approximately 1502.87, suggesting that the baseline price for an Airbnb listing (before accounting for the other variables) is around $1502.87. The coefficient for the number of reviews is slightly positive (0.019), but with a p-value of 0.434, it is not statistically significant at common significance levels. This implies that there is no strong evidence that the number of reviews has a substantial impact on the listing price. The 'minimum nights' variable has a coefficient of -1.624, which is statistically significant at the 0.05 level (p-value 0.040). This suggests that listings requiring a longer minimum stay are priced lower, perhaps to attract longer-term stays. The rating has a negative coefficient of approximately -2.09. This implies that as the rating increases by one unit, the price of the Airbnb listing is expected to decrease by about $2.105, holding all other variables constant. A potential explanation for this counterintuitive finding could be that listings with higher ratings are priced more competitively to attract guests and secure bookings, which in turn might lead to higher ratings due to better value for money. It is also possible that hosts with higher-rated listings aim to maintain such ratings by setting more reasonable prices.

However, the model's overall R-squared values are extremely low (0.0002 for multiple R-squared), indicating that the variables included explain almost none of the variability in Airbnb listing prices. The F-statistic has a p-value of 0.01985, suggesting that the model is statistically significant as a whole, yet its explanatory power is extremely limited. Therefore, while some coefficients are significant, other important factors not included in the model likely influence Airbnb prices. Future attempts to improve the model should explore the inclusion of additional relevant variables and the use of non-linear models that may capture the complexity of the data better.

Conclusion and Summary

In this exploratory analysis of the Airbnb Open Data for New York City, several key insights have emerged about factors influencing Airbnb listing prices. Geographic location plays a pivotal role. Contrary to our intuitions, listings in close proximity to downtown areas, such as Lower Manhattan, exhibited medium average prices, neither significantly low nor high when compared to more peripheral neighborhoods. This suggests a balanced market in these prime locations, where a high demand from tourists is met with a substantial supply of listings. Interestingly, the neighborhoods with the highest average prices were found in more suburban areas. A potential explanation is that these locations offer unique accommodations or luxury experiences in markets with less competition.

The analysis of minimum nights requirement revealed that listings with a 9-night minimum stay tend to have lower median prices, possibly as an incentive for longer-term stays. The relationship between the total number of reviews and listing prices showed varying local trends across different neighborhoods, indicating that the impact of reviews on pricing might be context-specific.

Moreover, our analysis discovered surprising trends related to the keywords used in listing titles. Listings featuring the keyword 'luxury' did not command higher prices as might be expected; instead, they exhibited lower median prices compared to other keywords. In contrast, listings with keywords such as 'village' or

'quiet' commanded the highest median prices, suggesting that these terms attract guests willing to pay a premium for specific experiences.

Our linear model, incorporating the number of reviews, minimum nights requirement, and customer ratings as predictors, indicated a statistically significant yet practically small effect of these variables on listing prices. Notably, higher customer ratings were associated with slightly lower prices, possibly reflecting a strategy by hosts to maintain high ratings through competitive pricing.

However, the model's low explanatory power underscores the complexity of the Airbnb pricing mechanism and suggests the need for a more nuanced model. Future research should consider incorporating a broader set of variables, such as property type and property construction year. Moreover, we can also employ more sophisticated modeling techniques to better capture the multifaceted nature of Airbnb pricing strategies. This analysis lays the groundwork for further investigation into the pricing landscape of Airbnb listings, offering a foundation for hosts, guests, and policymakers to better understand and navigate the sharing economy.