

New York’s Airbnb List Price Drivers

Group 12



Gan Jin

Denise Person

CAROLINA CARRILLO

Geuntae Jang

**MOTIVATION AND SUMMARY**

With Airbnb becoming a more popular choice of accommodation for travelers we started thinking what type of factors could potentially impact prices, be it number of rooms in a house, number of bathrooms, amenities or even location. Based on these questions we determined that the best way to see the impact was to analyze each Borough in New York City and compare how each factor affected price. In addition to these questions we also attempted to analyze whether rent prices in the city also had an impact on Airbnb, or vice versa. Unfortunately, due to limited data sources this analysis was done with a very small sample size compared to the actual number of properties available for rent in the city.  
  
In order to conduct our analysis we used different sources, InsideAirbnb, City of New York API, Streeteasy, and Zillow. Each of these sources provided different information, InsideAirbnb, Airbnb’s own API, was our main data source, we used it to pull information from current listings such as Airbnb type, prices, bedding types, number of bathrooms and more. From Streeteasy we pulled a small sample size of properties available for rent in the city, the details used for the analysis consisted on location, number of rooms and price. City of New York API was used to analyze the population density in each borough and compare it to the number of Airbnb’s in each borough.  
  
Moving on to the cleanup process, out first step was to clean up all files by removing rows without information or with NA and NULL values. After this we created a DataFrame leaving only the columns needed for the analysis. A few of the noticeable changes when cleaning up our Data sources consisted on keeping only 22 columns out of 106 from the listings file used for Airbnb exploration. Apart from this 10124 lines were dropped from this same file due to missing information, this was especially important when analyzing the impact of reviews as the columns with missing information were related to this variable. The same process was done for the other data sources; however the changes were not as significant as with Airbnb.

**DATA ANALYSIS**

Several variables were explored to determine if they influenced Airbnb list price. Two approaches were used in the analysis. The first approach involved creating various visuals including pie charts, bar charts, scatter plots, box plots and heat maps to see if there were any trends. In the second approach, a correlation matrix was created.

**Graphical Review of Visuals**

One of the first graphs we created was a histogram of the distribution of prices in New York City to eliminate outliers. The histogram is skewed to the right indicating that the prices are not normally distributed. All prices exceeding $1,000 (approximately 239) were excluded from the analysis. We believed that including prices up to $1,000 was reasonable due to the high prices in the Manhattan market. In addition, developers of luxury condominiums/rentals are reserving certain sections of new developments for Airbnb rentals. Because of this, we expect the distribution of prices to be multi-modal in the future.

A screenshot of a cell phone

Description automatically generated

***Variable: Location***

We expected that location would be a major contributor to Airbnb list prices in New York City with list prices of Airbnb rentals located in Manhattan being the highest. (Prices ranged from $10 to $10,000 for a Manhattan Airbnb rental.) Pandas were used to calculate the average price of each borough in the chart below, which was created using Matplotlib. Manhattan has the highest average price of all five boroughs at $197.00, followed by Brooklyn with $124.28, Staten Island with $115.68 (initially a surprise to us), Queens with $98.37 and the Bronx at $86.27. For the most part, the results were consistent with expectations as Manhattan is a tourist destination and neighborhoods in Brooklyn are close in proximity to Manhattan. We were initially surprised by the fact that the price of Airbnb rentals in Staten Island was close to that in Brooklyn, but noted that Staten Island had a small number of listing, 352, compared to 19,809 listings in Brooklyn.

A screenshot of a cell phone

Description automatically generated.

A box chart of each borough was also created and revealed similar results.

A picture containing screenshot

Description automatically generated

In order to further validate our analysis, we took a more granular look at the data and created a bar chart below showing the number of listings in each borough at list prices increasing in increments of $50. In price bands exceeding $100, there are more Airbnb listings in Manhattan than in any other borough.

A screenshot of a cell phone

Description automatically generated

A heatmap of NYC zip codes was created using GeoPandas. The heatmap comparing the average price was less revealing at first glance. Zip code 10309 (Richmond County) in Staten Island and 11360 (Bayside) in Queens had two of the highest average prices in New York City. However, both were based on a smaller number of listings, which explained this when taking an average. The zip code in Manhattan with the highest price was 10281 (World Trade Center).

A close up of a logo

Description automatically generated

***Variable: Room Type***

There were three different room types in the Inside Airbnb data that was downloaded: entire home/apt, private room, shared room. There were very few shared rooms—1,143. Entire home/apt and private rooms represented most of the listings at 25,062 and 22,028, respectively. The average price of the entire home/apt room type at $212.45 was higher than that of the private room and share room, which were $89.47 and $68.49, respectively.

A screenshot of a cell phone

Description automatically generated

We also created a bar chart below showing the number of listings for each room type at list prices increasing in increments of $50. Except for the entire home/apt Airbnb rental with a list price below $100, list prices for entire home/apt rentals exceeded those of the private room and shared room rentals. Note that Manhattan has more Airbnb listings than any borough for list prices in bands exceeding $100. This further supports our expectation that list prices for the entire home/apt room type exceeds private room and shared room list prices.

A screenshot of a cell phone

Description automatically generated

***Variable: Number of Guests Accommodated***

We examined the relationship between the number of guests an Airbnb rental can accommodate and price by creating a Box Plot with the number of guests on the x-axis and the list price on the y axis. Generally, the box plots followed a consistent trend. The more guests an Airbnb rental could accommodate, the higher the median list price. There were a few inconsistencies such as the low price of the box plot of 13 guests that could be accommodated compared to the other box plots with higher number of guests that could be accommodated which was due to a small number of data points. Other factors contributing to the inconsistencies are related to other variables.

A close up of a logo

Description automatically generated

***Variable: Number of Bedrooms***

Box plots were created to show the relationship between price and the number of bedrooms. The box plot showed a consistent trend with the more bedrooms in an Airbnb rental, the higher the prices. As in earlier box plots, there were a few inconsistencies due to the small number of data points in that band and other variables.

A screenshot of a cell phone

Description automatically generated

***Variable: Number of Beds***

The relationship between price and the number of beds is shown in the box plots below. The plot shows that the more beds in an Airbnb rental, the higher the list price. There were a few inconsistencies due to the small number of data points particularly for box plots with high numbers of beds. We also noted that the box plot of 13 beds appears to not have a median. In this case, there were few data points. Most of the data points in that band were equal. This resulted in the median equaling the 75th quartile.

A screenshot of a cell phone

Description automatically generated

***Variable: Number of Bathrooms***

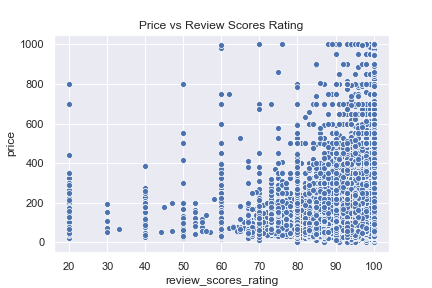
The box plot showing the relationship between price and the number of bathrooms follows a consistent trend where the higher the number of bathrooms in a rental, the higher the price. The box plot of 5 bathrooms appears to have no median. However, the median is the 75th quartile. There are only a few data points in this box plot interval and most of them have a price of $800, which also happens to be the 75th quartile.

A picture containing screenshot

Description automatically generated

***Variable: Review Score Rating***

A scatter plot was created to show the relationship between price and review score ratings. Most of the review score ratings were 70% or higher. There does appear to be a trend where the higher the score, the higher the list price.



**Correlation Matrix**

A correlation matrix was created to show the relationship between two variables. Pandas dataframe.corr() was used to find the correlation of all the columns of the data.

MinMaxScaler of the skLearn.preprocessing package was used to transform the data into values between 0 and 1 in order to use the Pandas dataframe.corr() function. Dataframe.corr() produced a table of correlations between each column, which was then used to as inputs to the Python Seaborn library to create the correlation matrix below.

A screenshot of a cell phone

Description automatically generated

The table of correlations corresponding to the correlation matrix above is too exhaustive to show in the report. However, we have saved them in a csv file named correlation.csv and created a summary table below which shows those variables correlated to price that exceed 10%.

As can be seen in the correlation matrix, there are no variables that are strongly correlated to price. The most correlated variable to price is the number of guests accommodated at 54.9%. The next highest correlation is the variable entire home/apt, followed by number of beds, number of bedrooms, and number of baths. We would expect that these variables are all highly correlated. An entire house would have more bedrooms, beds and baths and could therefore accommodate more guests.

Private room is negatively correlated with price with a coefficient of -46%. As price increases, we would expect fewer residents to rent a private room. The variable shared room is also negatively correlated with price. However, the number of shared rooms was small relative to the number of private rooms to be significant. Longitude was also negatively correlated with price. We noticed that the more west the longitude was, the higher the price. We did not use the boroughs as a variable, which we expect would have shown a stronger correlation between price and borough.



**DISCUSSION/CONCLUSION**

Our goal was to determine if certain variables influenced price. We did this by graphing and plotting charts, which showed relationships between price and a variable, and creating a correlation matrix. Although the correlation matrix did not show strong correlations with price, we can conclude that with the correlations shown and the visuals used in the analysis that location, the room type, the number of bedrooms, the number of beds, the number of bathrooms and review scores do influence price.

**POST MORTEM**

The scope of our proposal which in the beginning was seemingly narrow turned out to be broader than anticipated. This is because there are many variables that influence Airbnb list price. In our research, we realized that some factors are endogenous such as number of rooms, number of bathrooms and location and others are exogenous such as housing/apartment rental prices, hotel prices, tourism. Although we focused our analysis on the endogenous variables, the challenge is that the endogenous and exogenous variables are all correlated to some degree and the endogenous variables are also correlated with each other.

If we had more time, we would perform regression analysis. However, this would first involve us exploring the pairwise relationships on the variables in the correlation matrix (other than price). For those that are highly correlated or predictive of each other, we would likely remove them to remove redundancies and thus avoid skewing the data.

Another analysis we would consider, is to isolate the data set based on the variable analyzed. For example, if we wanted to analyze how location influenced price, it would be helpful to select one room type say entire house/apartment, two bedrooms and two bathrooms, only, creating multiple combinations of the variables in the data set. This would isolate the influence that other variables have on price and more likely provide a more accurate representation of how location influences price.