Regression Analysis Project Report

Finding What Determines the Price of an Airbnb in New York City

Sofia Marquez, Valeria Rodriguez, Liam Tormey, Luis Carlos Ahuage, Gunner Harvey

University of San Diego

December 2021

Abstract

In this report, we will use multiple regression analysis to find what best determines the price of an Airbnb in New York City. We will evaluate multiple variables such as Price, Location, Area, Minimum Number of Nights, Availability and listing space type. Then we used the estimated regression model to predict the price for an Airbnb. We find that the location, type of room, availability, and number of reviews are important factors in determining the value for an Airbnb in New York City.

Table of contents

Introduction

When trying to decide on what Airbnb is the best option for your trip, it can be overwhelming choosing between so many different accommodations and prices. This is why it is important for customers to be able to determine what factors contribute to the pricing of their Airbnb options, in order to be able to decide on the best choice. In turn, this also helps Airbnb renters who can better decide on properties to rent by understanding the factors that will have the greatest impact on their potential rent price.

This study is focused on discovering what factors contribute most to the constantly varying prices of an Airbnb. To measure this, we used a data set of Airbnb prices in New York City. The price is based on the neighborhood, number of reviews, whether they are renting the entire home or just a room, the minimum nights for the stay, and the availability of the Airbnb out of the 365 days out of the year.

Using OLS to estimate the regression line that best describes the relationship between value score and independent variables, we found that the minimum number of nights is not statistically significant to the price, but neighborhood, number of reviews, size (entire home or room), and number of days available, are.

Data

The data used in the analysis comes from Dgomonov, a Kaggle contributor from Drexel University, who gathered it from the website/source *Inside Airbnb*. The data set is called "New York City Airbnb Open Data Airbnb Listings and Metrics," and is determined by the Airbnb status in New York City, New York in the year 2019. The data focuses on Airbnb's in neighborhoods in New York City, and provides information on the location, area, availability, number of reviews, and if the airbnb is for the entire home, or just a specific room or space.

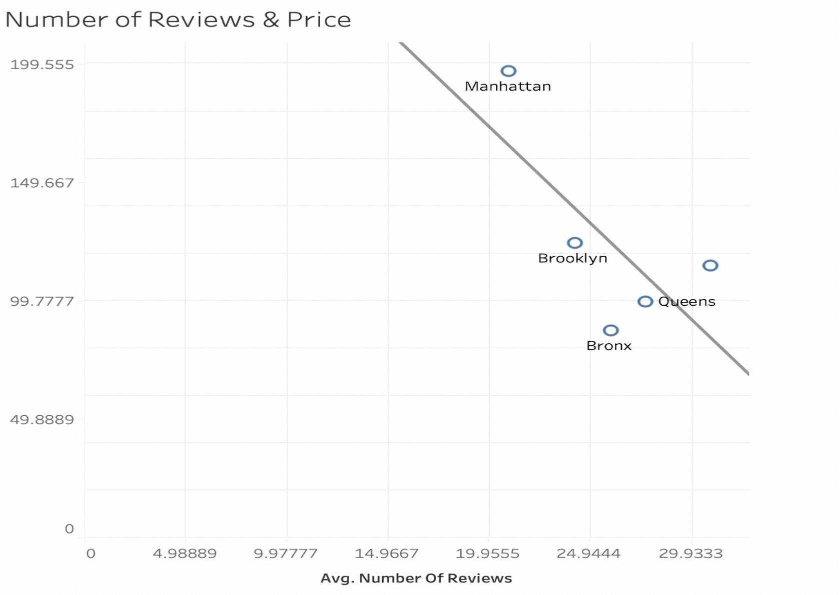
Below is Table 1, the summary statistics for the variables in the dataset. From the data, it is shown that the mean price of an Airbnb listing in New York City is $150.09. As well as, the price of an Airbnb listing ranges from $0 to $10,000, with a standard deviation of $238.33. From this it is clear that there is a large variation in price.

Table #1- Summary Statistics

	Price	*Entire Home*	*Minimum # of nights*	*Availability/ 365*	*# of reviews*	*Brooklyn*	*Manhattan*	*Other*
Mean	150.093	0.517	7.083	108.115	25.265	0.418	0.441	0.141
Std. Error	1.126	0.002	0.099	0.617	0.217	0.002	0.002	0.002
Median	105.000	1.000	3.000	35.000	7.000	0.000	0.000	0.000
Mode	150.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000
Std.Deviation	238.326	0.500	20.973	130.604	46.008	0.493	0.497	0.348
Minimum	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000
Maximum	10000.000	1.000	1250.000	365.000	629.000	1.000	1.000	1.000
Count	44828	44828	44828	44828	44828	44828	44828	44828

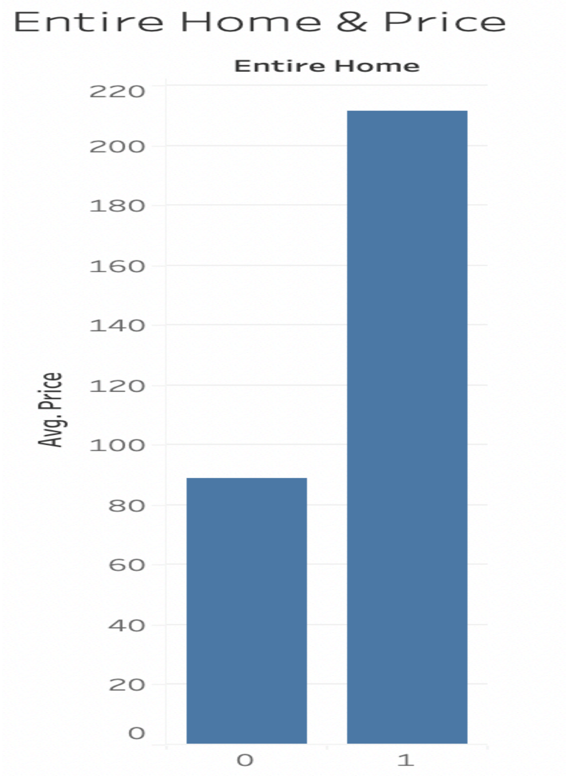
In Figure 1, “Number of Reviews and Price”, it shows that as the number of reviews increase, the price will decrease. Our reasoning behind this is most people who write reviews tend to write negative reviews. Meaning that when people look at reviews on the Airbnb website, they tend to see negative reviews. This causes less people to lease these Airbnb’s, which causes the rent price of the Airbnb to decrease. The figure also shows that no matter the number of reviews, Manhattan still has the highest prices, followed by Brooklyn. We predict that the number of reviews will have a negative impact on the price due to the majority of the points being below the trendline.

Figure #1-Number of Reviews & Price



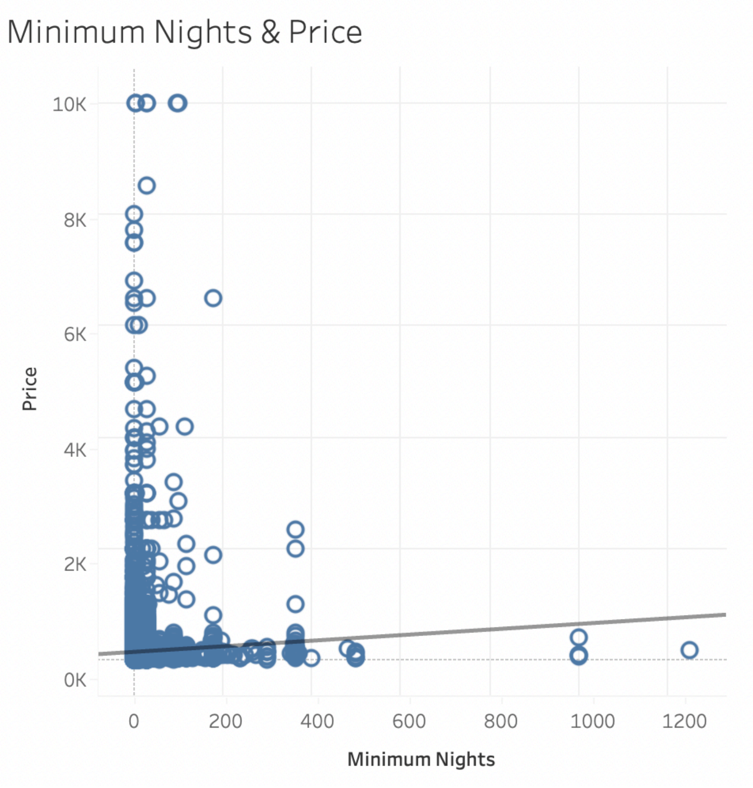
In Figure 2, labeled “Entire Home & Price", we are expecting that this data would be statistically significant. The bar graph is showing whether the Airbnb rented is for the entire home, or just a room/shared room. Those with a value of 1, represent Airbnbs that include the entire home, whereas those with 0 represent a single room. We can see that if the entire house is being rented then it would cost around $210, which is nearly triple what a room would cost.

Figure #2-Entire Home & Price



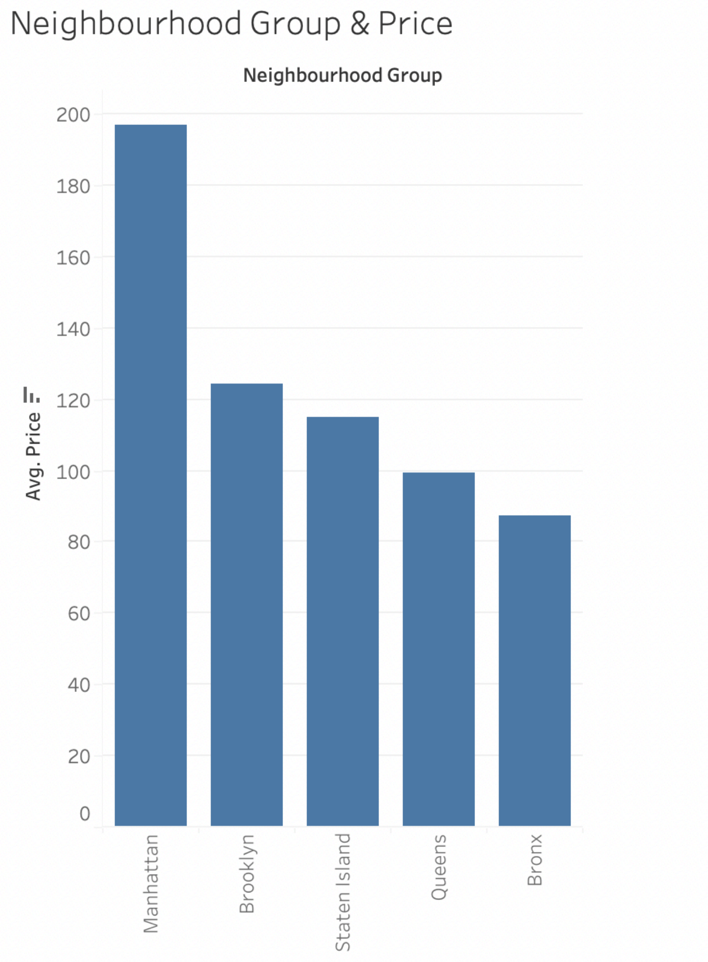
In Figure 3 below, “Minimum Nights and Prices”, we are expecting that the number of minimum nights will be insignificant to the price of rent. We can tell this by analysing the inconsistency of the data, as there are significant outliers, and no clear correlation on the scatterplot. The trendline helps show this, as there also isn’t any clear correlation with the line.

Figure #3- Minimum Nights & Price



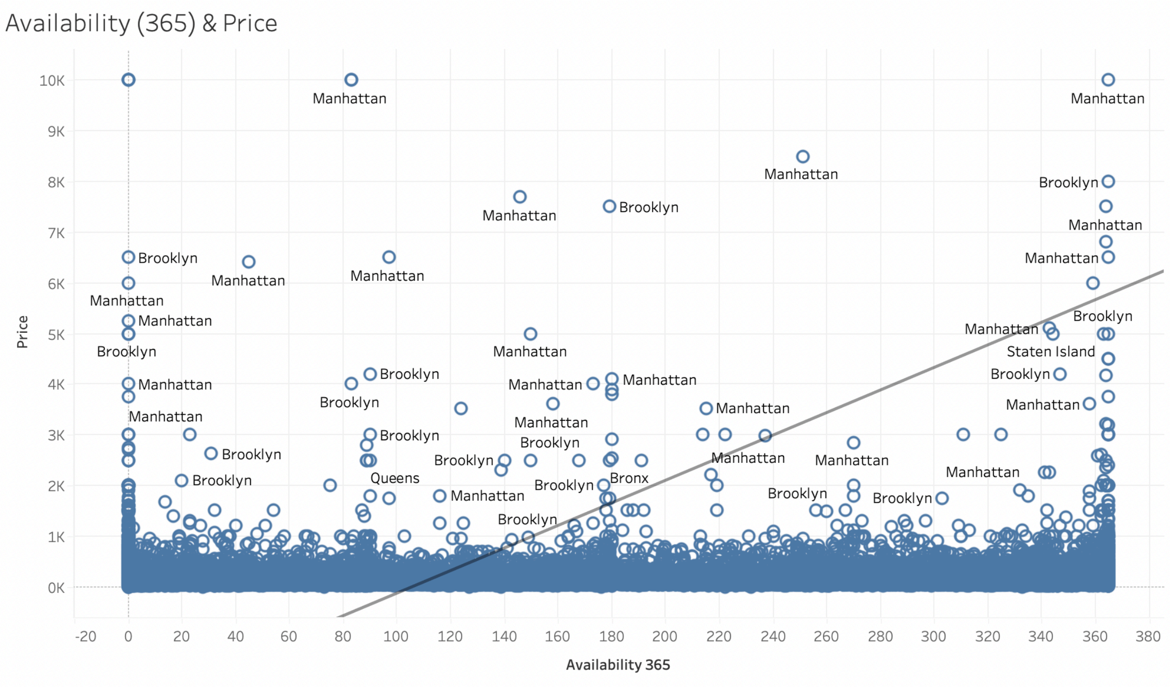
In Figure 4 below, labeled "Neighbourhood Group & Price", we are looking at the areas in New York City along with their price. Our group expected that this would be statistically significant. As shown below, Airbnb's in Manhattan are most expensive, and Brooklyn comes in second, which would make sense as they are the more prominent/desired areas in New York. The average price in Manhattan would be around $200 a night and Brooklyn's average would be around $120 a night. The other neighborhoods, which include Staten Island, Queens, and the Bronx, will have a lower price compared to the two others, with all of their averages being less than $120. Because of this, we decided to group those three neighborhoods and labeled them as "other".

Figure #4- Neighbourhood Group & Price



In Figure 5 below, “Availability (365) and Price”, it shows the correlation between the price and the number of days that Airbnb listing is available throughout the entire 365 days in a year. There is a positive trendline associated between these two variables. Therefore, our prediction is that it will have a positive impact on price. Potential reasoning behind this is that there are certain times of the year that prices are higher due to high seasons, when many people travel, for example, Christmas and other holidays.

Figure #5- Availability (365) & Price



Since there was multiple categorical variables, the following dummy variables have been formulated:

Entire Home − Room Type = { 1 if the room type is entire home 0 *otherwise}*

For organizing the neighborhood groups we compared Manhattan and Brooklyn to rest, due to those correlating with the highest price.

Manhattan − Neighborhood Group= { 1 *if* ***the listing is in Brooklyn 0 if listing is otherwise}***
Brooklyn − Neighborhood Group= { 1 *if* ***the listing is in Manhattan 0 if listing is otherwise}***

Next, we will combine this data to create an estimation strategy utilizing these variables.

Estimation Strategy

First, we set **price** as the dependent variable, then developed an estimated regression equation using Minimum Number of Nights, Availability Out of the Year, Number of Reviews, Entire Home, Manhattan, Brooklyn as the independent variables. However, we found the Minimum Number of Nights, as predicted, to be statistically insignificant and found the estimated coefficient to be essentially zero. Because of this, we decided to drop the Minimum Number of

Nights from the set of independent variables in our model. The regression model below, shows the model used to describe the determinants of Price.

$$\textit{Price} = \beta 0 + \beta 1\ \textit{Availability/365} + \beta 2\ \textit{Number of Reviews} + \beta 3\ \textit{Entire Home} + \beta 4\ \textit{Manhattan} + \beta 5\ \textit{Brooklyn} + \varepsilon$$

Then we created the regression estimated regression equation to predict the straight-line relationship between price (dependent variable) and the independent variables. We expect to see a slight positive relationship between price and availability out of the year. We also expect to see a slight negative relationship between price and number of reviews. In regards to the entire home variable, we expect Airbnbs that rent out the entire home to have a higher price then units that only rent out a room. As for location, we expect Manhattan to have a higher price than Brooklyn. We also expect Manhattan to have a higher price than all the other locations.

Next, we will introduce and analyze the results of the regression analysis. We will also explain how each independent variable in our regression model affects the dependent variable, price. As well as, if the predicted outcomes were correct.

Regression Analysis Results

Below is the estimated regression line with the values coefficients for price and the independent variables inserted in:

$$\textit{Price} = 38.227 + 0.179\ \textit{Availability/365}\ \text{--}\ .303\ \textit{Number of Reviews} + 112.530\ \textit{Entire Home} + 76.652\ \textit{Manhattan} + 22.230\ \textit{Brooklyn}$$

When analyzing the data in Table 2, the main concern was the R Square value of 0.092. This means that the estimated model only explains 9.2% of the variation in regards to the dependent variable, price. So with this data, there are a lot of other variables that impact the price of Airbnbs that aren't accounted for.

Table #2- Regression Statistics

Multiple R	0.303
R Square	0.092
Adjusted R Square	0.092
Standard Error	228.853
Observations	48895

However, based on the F-test in the ANOVA results (Table 3), there is still evidence that the data is statistically significant. In table 3 below, the F statistic is 825.695, which with a p-value of 0.000, is indeed significant at 1 percent significance level.

Table #3- Regression Results

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	259468227.148	43244704.525	825.695	0
Residual	48888	2560445563.277	52373.702		
Total	48894	2819913790.425			

Dependent Variable: PrIce	*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	38.228	3.096	12.349	0.000	32.160	44.295	32.160	44.295
Entire Home	112.530	2.108	53.382	0.000	108.398	116.662	108.398	116.662
Availability /365	0.179	0.008	21.904	0.000	0.163	0.195	0.163	0.195
Number of Reviews	-0.303	0.024	-12.748	0.000	-0.349	-0.256	-0.349	-0.256
Brooklyn	22.230	3.189	6.972	0.000	15.980	28.480	15.980	28.480
Manhattan	76.653	3.186	24.058	0.000	70.408	82.898	70.408	82.898

In the second part of Table 3 above, we mainly focused on the relationship between the independent variables and the P-value. It is evident that the dummy variable, Entire Home, is statistically significant, with a coefficient of 112.530, at a ninety-five percent significance level since the p-value is 0.000. This means that on average an Airbnb that offers the entire home/apartment instead of a single room, price increases by 112.530 units. As for availability out of the year, it has a coefficient of 0.179 and a p-value of 0.000 it is statistically significant, as well. So for every unit increase in availability, there is a 0.179 unit increase in price. In regards to the number of reviews, with a coefficient of -0.303 and a p-value of 0.000, it is also statistically significant. However, unlike the other independent variables, the number of reviews causes a decrease in price. For every increase in units of reviews, there is -0.303 decrease in units of price. Lastly, the two dummy variables for Brooklyn and Manhattan, are both statistically significant (both have p-values of 0.000). So for every time the Airbnb is in Brooklyn (vs. not) there is a unit increase of 22.230 in price and for every time there is one in Manhattan there is a 76.653 increase in units of price.

Discussions and Limitations

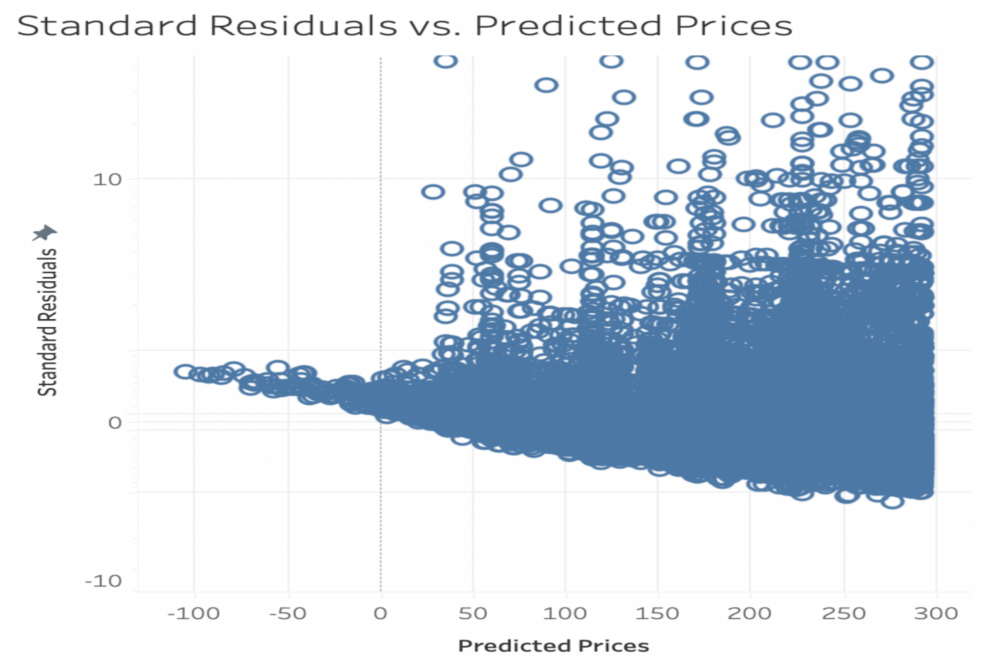
First and foremost, the model was successful in substantiating our expectations of the coefficients. All had the expected signs, and there was a clear substantial increase in price with “entire home” and “Manhattan.” However, we did not predict that “entire home” would be significantly higher than all the other coefficients. Which is helpful for Airbnb sellers to know that if they want to maximize the value of their rentals, that it is important to rent out the whole property.

Although the coefficients were valuable, the regression results have shown that the independent variables contribute little to the overall variation of the dependent variable (price). This is likely due to Airbnbs having a lot more variables in pricing than the individual variables tested. This model could be greatly improved with more variables such as dates the home was rented, square footage, cost estimate of home, to name a few. If we were to repeat this study, we would try to gather this data with the goal of analyzing a much larger amount of variation for the Price. Below, we will perform a residual analysis, to assess the model assumptions and the credibility of the previous results.

Residual Analysis

Figure 6 below, shows the relationship between standard residuals and predicted prices. The range for the standard residuals, (between -10 and +10) does not contain all of the data points. This shows that there is a significant amount of outliers present in the data. Also it is clear that the spread of the residuals are not equal at each predicted price level. Therefore, we can also concur that the assumption of constant variance is not met. Based on these results, the model is not a strong representation of the relationship between the dependent/independent variables.

Figure #6- Standard Residuals vs, Predicted Values



Additionally, we ran the normal probability plot for the model (as shown below, Figure 7). In this there are apparent outliers that skew the data, so it is clear that the normality assumption is not met.

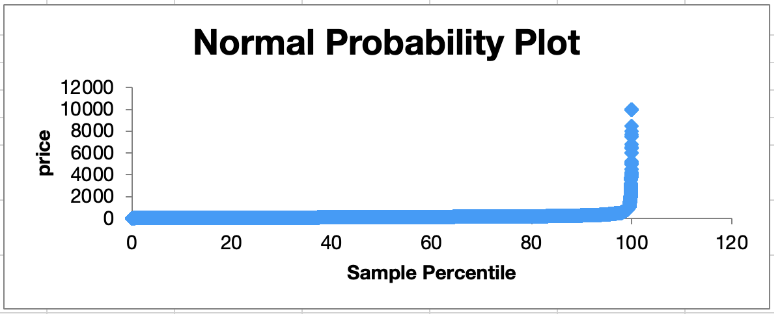


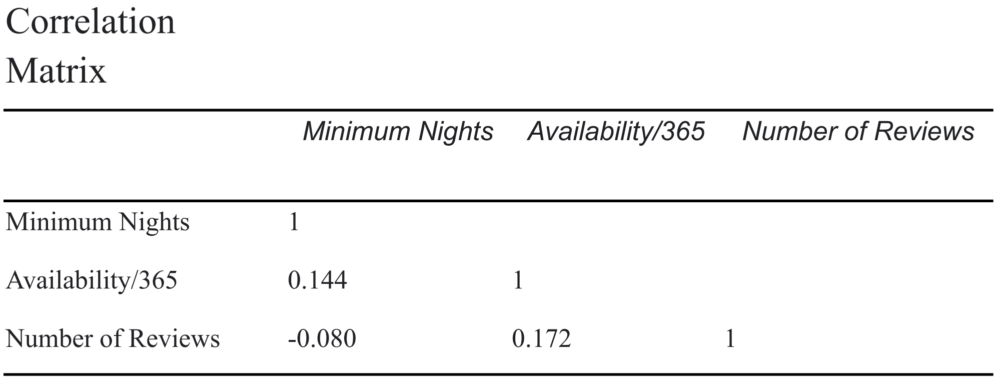
Figure #7- Normal Probability Plot

Overall, the residual analysis we conducted proved that a large amount of variation in the dependent variable is not due to the independent variables and the estimated regression line is not necessarily a strong representation of the data.

Multicollinearity

Below is a correlation matrix (Table 5) for the quantitative variables in the dataset. Based on the chart, it is clear that the relationship between each independent variable has low/negative correlation. This is important, as it shows that multicollinearity is not a problem at all. Therefore, it further proves the regression results are trustworthy.

Table #4-
Correlation Matrix



Correlation Matrix

	Minimum Nights	Availability/365	Number of Reviews
Minimum Nights	1		
Availability/365	0.144	1	
Number of Reviews	-0.080	0.172	1

Model's Predictive Ability

Overall, based on the overall residuals gathered from the model, it is unlikely that it will be very effective at predicting a single Airbnb price. However, it is still important to test the model's predictive ability by assessing how it performs with a single data point, as seen below.

Table #5- Observation Number 1 in the Data

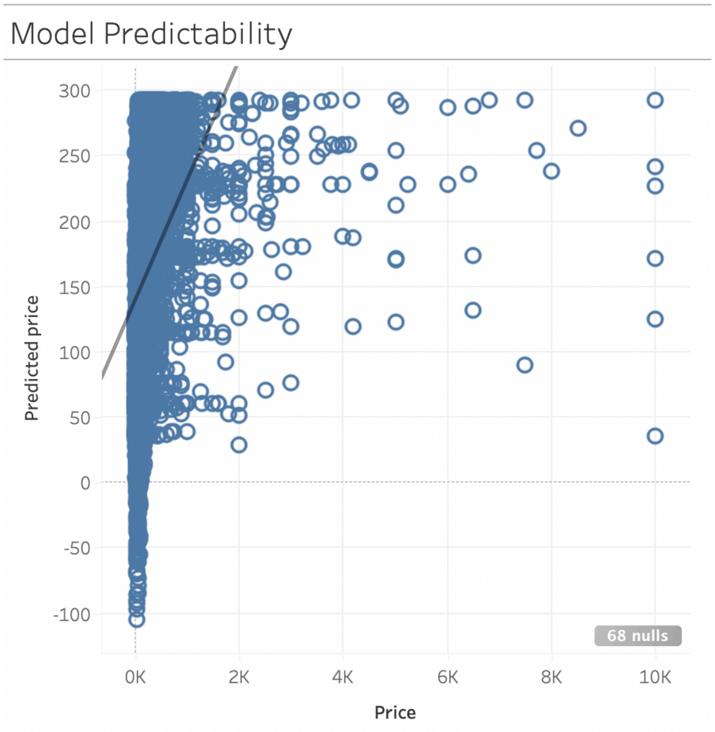
Price	Entire Home	Minimum Nights	Availability/365	Number of Reviews	Brooklyn	Manhattan
149	0	1	365	9	1	0

Price = 38.227 + 0.179 X 365 (Availability/365) -- .303 x 9 (Number of Reviews) + 112.530 X 0 (Entire Home) + 76.652 X 0 (Manhattan) + 22.230 X 1 (Brooklyn)

The Price predicted for this particular observation is 123.065, whereas, the actual price is 149. which means the error term is 149 - 123.065= 25.985. Although 25.985 is a low error term, we infer that this was just a matter of luck, as the majority of the other residuals are much higher.

Below in Figure 8, we have built a scatter plot that compares the predicted Price to observed Price which shows that the model doesn't do a reliable job of predicting the dependent variable.

Figure #8- Model Predictability



Conclusions and Recommendations

In this study, we used data on Airbnb listings in different neighborhoods in New York City. When trying to determine which Airbnb to rent, the price is determined by numerous variables such as the rent space (entire home), the number of reviews, the availability, and the area/neighborhood. The minimum number of nights showed no statistical significance when showing if the listing would be more expensive. We used price as the dependent variable in a multiple regression and tested to see if availability out of the 365 days in a year, number of reviews, and minimum nights are good determinants of the price of each listing. We found that the location and type of room (or entire home) is what most assists in determining the price. This proved to be the most statistically significant which would make sense considering the location and amount of space often influence the price. However, our r-squared value was extremely low, so the independent variables did not have a huge impact on the variation of our dependent variable, price.

For a potential business model, if we had more independent variables, the study would be able to help customers accurately find the best Airbnb listing based on the price increase for each variable they prefer. In turn, if the r-squared was larger, a business could use this data to help customers find an Airbnb focused on their specific needs for the lowest price. Such as if a customer specifically cared about location, they would be able to see how much that would likely impact price. However, based on the findings from this study, what we can establish for Airbnb customers/renters, is that location and if the entire home is for rent, have the greatest impact on price.

References

Dgomonov. “New York City Airbnb Open Data.” *Kaggle*, 12 Aug. 2019, https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data.