Analyzing New York City Airbnb Data Using Data Visualizations and Machine Learning Techniques

Hua Wei

9/26/2019

Introduction

In the last few years, Airbnb has become an increasingly popular marketplace for arranging or offering lodging services. It provides more and more opportunities for visitors to explore the destinations and choose their desired lodging services. The millions of listings posted on Airbnb every day is an indication of that. Analysis of the listings can provide insights for making business decisions, understanding customers' and providers' needs and behaviors, creating marketing campaigns, enhancing existing services, and implementing new services.

This study is intended to look at the New York City Airbnb Open Data and investigate characteristics of the listings. Specifically, it is interesting to find out whether room price in the listings can be predicted based on other information such as number of nearby venues, room type, and number of reviews, and how good the prediction is.

Results of the study will be particularly meaningful to providers (or hosts) who are keen to know what factors influence price the most. In addition, they may want to use the study results to come up with room rates that are reasonable and to their advantage.

Data

In this study, I used the New York City Airbnb Open Data posted on https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data. It is in the form of a CSV file. The dataset describes the listing activity and metrics in NYC, NY for 2019. It includes over 40,000 rows and 16 columns as follows:

Variable	Notes
id	listing ID
name	name of the listing
host_id	host ID
host_name	name of the host
neighborhood_group	borough
neighborhood	area
latitude	latitude coordinates
longitude	longitude coordinates
room_type	listing space type
price	price in dollars
minimum_nights	amount of nights minimum
number_of_reviews	number of reviews
last_review	latest review
reviews_per_month	number of reviews per month
calculated_host_listings_count	amount of listing per host
availability_365	number of days when listing is available for booking

In addition, I used the dataset that contains the 5 boroughs in New York City and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood. The dataset is readily available at the website (https://geo.nyu.edu/catalog/nyu_2451_34572). I also used the Foursquare location data which I obtained through API calls.

Methodology

After the dataset was read in, I checked its size and the type of each variable. Then I checked the number of missing values for each variable. I decided to get rid of four variables, namely, id, host_name, last_review, reviews_per_month, because they had the most missing values and would not be used for the following analysis. Then I obtained the descriptive statistics for the remaining variables to get a sense of what the average value, range of values, minimum, and maximum would be for each variable.

To understand the data better, I created a bar chart that compares the number of listings across boroughs, and another bar chart that compares the number of listings between room types and across boroughs. Because price was what I wanted to make predictions about, I created a boxplot to display the distribution of price within each borough. All my subsequent data explorations and visualizations were focused on price, how much it varies, and what may be the factors that influence it.

To better predict price, I decided to utilize the Foursquare API to find out the number of nearby venues in the neighborhood where the listed room (or home) was located. This might be a factor that influences the price of the listing. To make the API calls, I used the above-mentioned dataset that contains the boroughs, neighborhoods, and latitude and longitude coordinates of all neighborhoods. Based on the observation that there are over 20,000 listings in Manhattan in the original data file, I thought that only Manhattan data would be sufficient for my subsequent analysis. Therefore, I made API calls to explore the neighborhoods in Manhattan only and created a new data frame based on the results of the API calls. The new data frame includes the name of the neighborhood and the number of nearby venues (variable named "number_of_nearby_venues") in that neighborhood. I then merged this data frame with the Manhattan data that has all the Manhattan listings. I got rid of listings with a price over 400, which was close to the "maximum" price for Manhattan as shown in the boxplot. Any price above that number could be considered as an outlier.

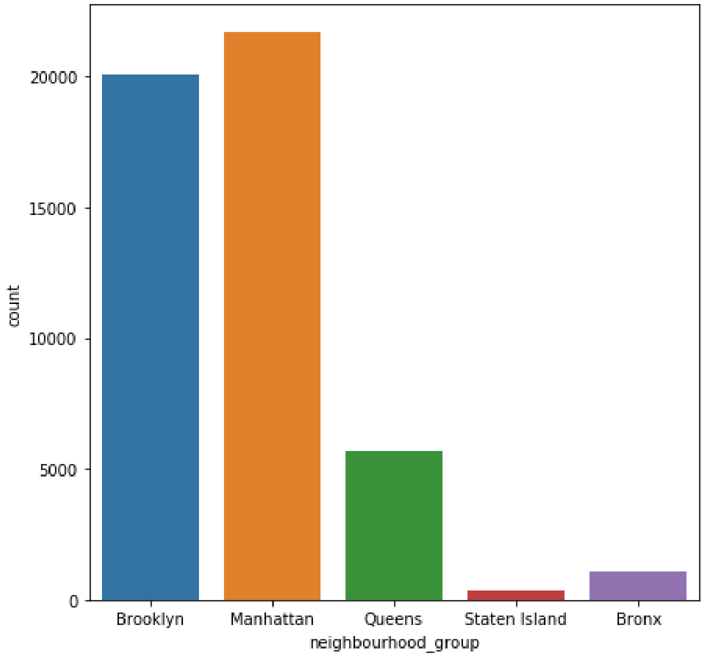
To make predictions about price, I decided to use the following variables: room_type, number_of_reviews, calculated_host_listings_count, minimum_nights, and number_of_nearby_venues. To further explore the data, I obtained a correlation matrix that shows the correlation between every pair of numerical variables in the data. I also generated scatterplots that show how price covaries with each of the five predictor variables.

To answer the research question, I used the multiple linear regression model to predict price with five predictor variables. The library "scikit learn" was imported to implement machine learning techniques on the data. Dummy variables were created on the variable "room_type", which was the only categorical variable, and the Manhattan data was split into training and test data. Training data were used to train the model, and test data to evaluate how well the model performed.

Results

Figure 1 shows the count of listings in each borough.

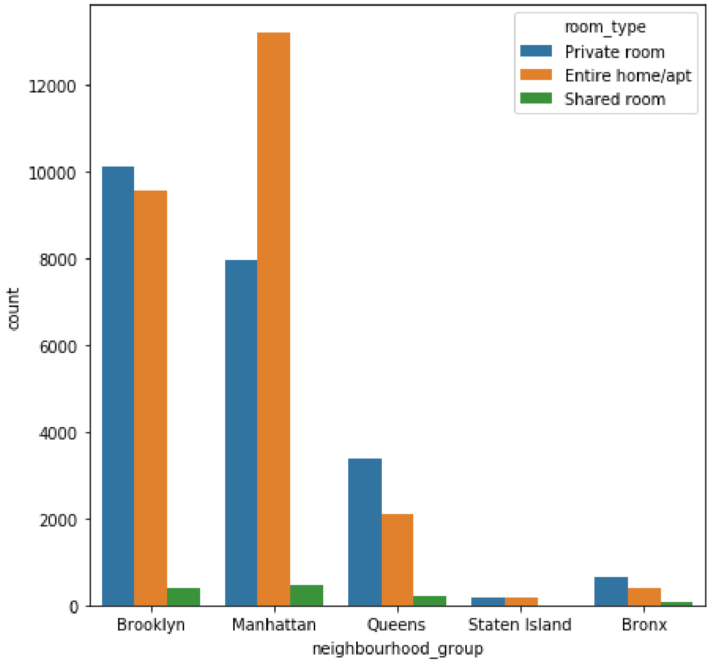
Figure 1. Count of Listings by Borough



As clearly shown, Manhattan has over 20,000 listings, and is the borough with the largest number of listings.

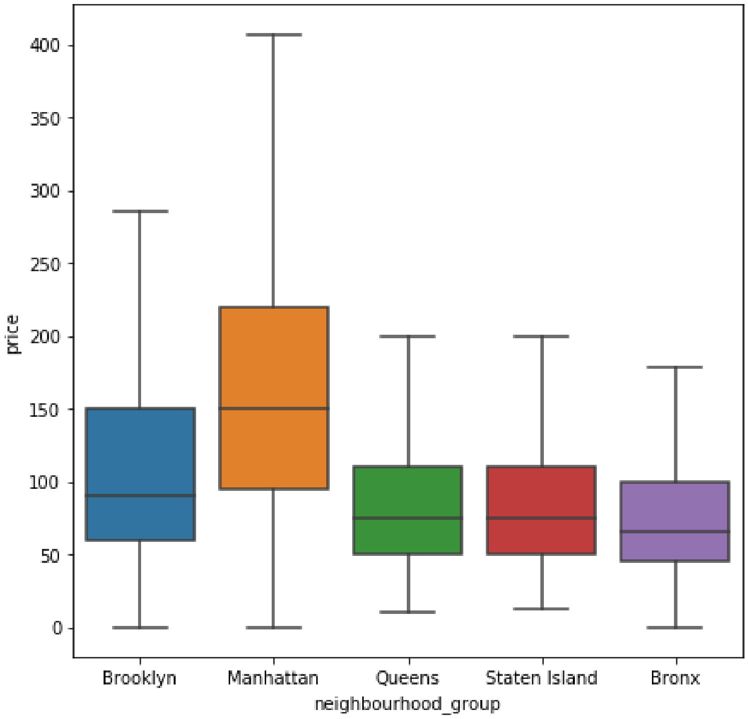
Figure 2 shows the count of listings by room type and borough.

Figure 2. Count of Listings by Room Type and Borough



As can be seen, most listings are either private room or entire home. Very few shared rooms are found in any borough.

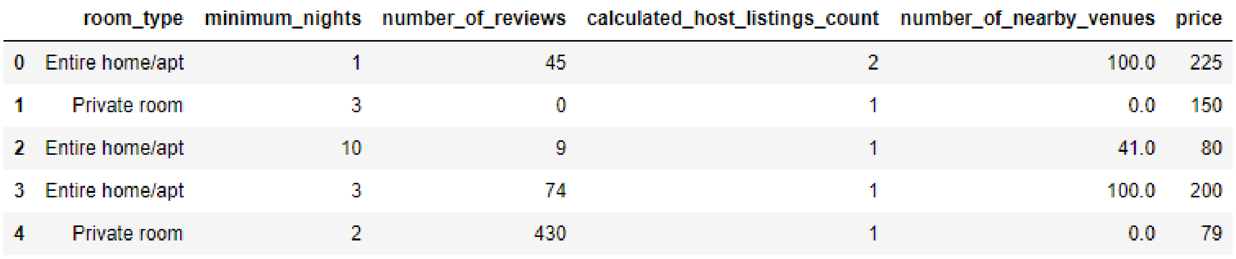
Figure 3 shows the distribution of price in each borough.



Outliers are not shown in the boxplots. The variation of price is the largest among listings in Manhattan.

Instead of building a model using all the listing data in New York City, I decided to focus on Manhattan because it has a large enough sample and may result in findings that are easier to understand and interpret than if I looked at the entire city. By joining the original Airbnb listing data with the newly created data from the results of the Foursquare API calls, I got a data frame that includes 20,366 rows and 6 columns. Table 1 shows what the data frame looks like. In this data set, price is the outcome variable, and the rest are predictor variables.

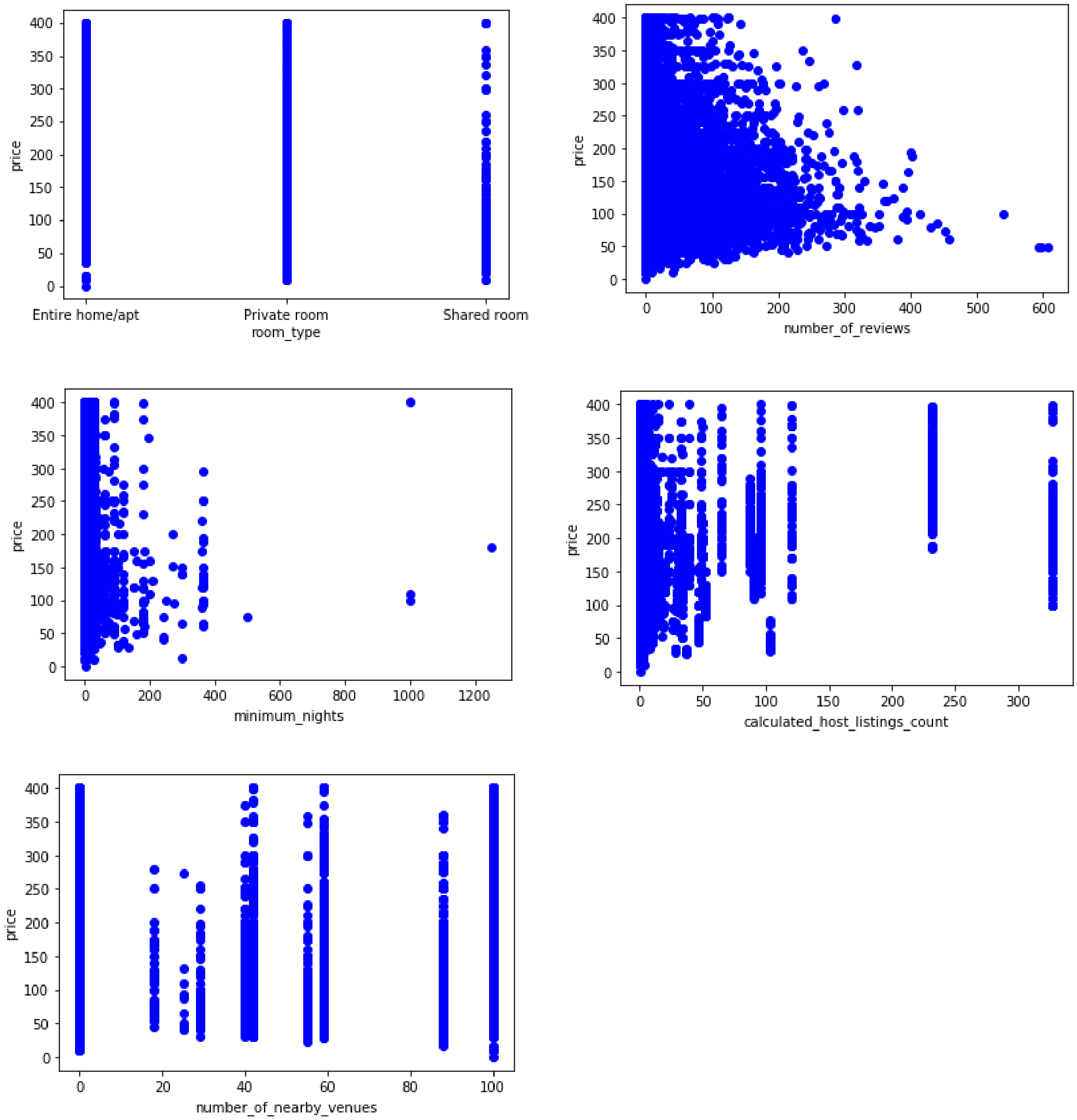
Table 1. First Five Rows of the Final Data Frame



	room_type	minimum_nights	number_of_reviews	calculated_host_listings_count	number_of_nearby_venues	price
0	Entire home/apt	1	45	2	100.0	225
1	Private room	3	0	1	0.0	150
2	Entire home/apt	10	9	1	41.0	80
3	Entire home/apt	3	74	1	100.0	200
4	Private room	2	430	1	0.0	79

The scatterplots between price and each of the predictor variables are shown in Figure 4.

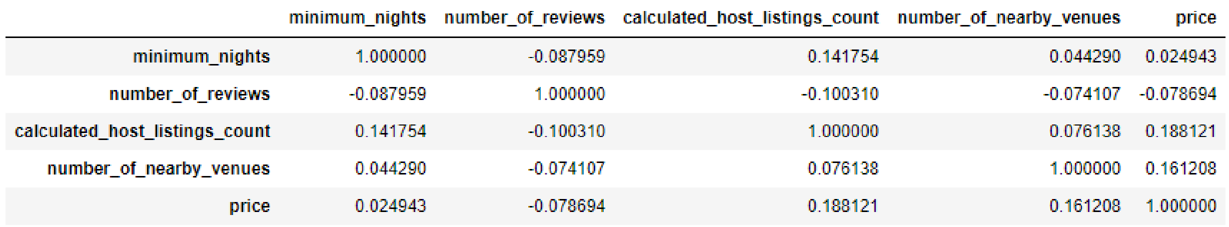
Figure 4. Scatterplots between Price and its Predictors



Not a strong trend is found in any of the scatterplots. In other words, price does not seem to be strongly correlated with any of the five predictors.

The correlation matrix between each pair of numerical variables is shown in Table 2. “room_type” is a categorical variable and excluded from the correlation analysis.

Table 2. Correlation Matrix between Price and its Predictors

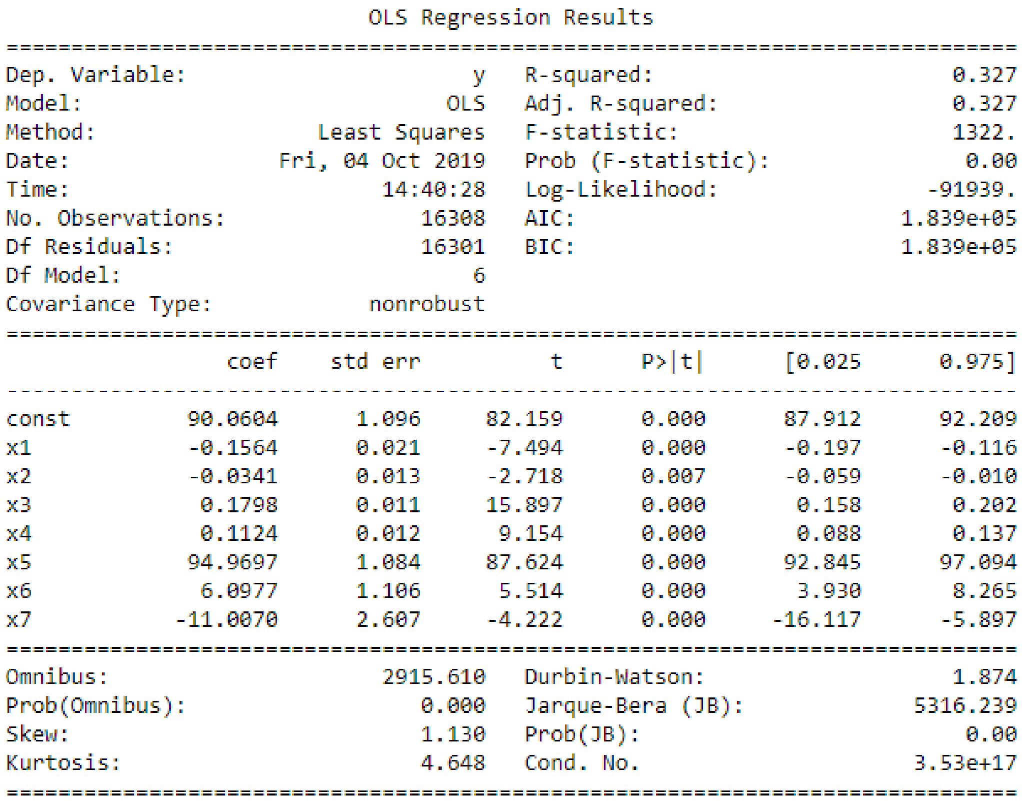


	minimum_nights	number_of_reviews	calculated_host_listings_count	number_of_nearby_venues	price
minimum_nights	1.000000	-0.087959	0.141754	0.044290	0.024943
number_of_reviews	-0.087959	1.000000	-0.100310	-0.074107	-0.078694
calculated_host_listings_count	0.141754	-0.100310	1.000000	0.076138	0.188121
number_of_nearby_venues	0.044290	-0.074107	0.076138	1.000000	0.161208
price	0.024943	-0.078694	0.188121	0.161208	1.000000

As shown, the correlation between price and minimum_nights is close to 0, which means no correlation. Price is negatively correlated with number_of_reviews, and the correlation is minimal. Price is positively correlated with calculated_host_listings_count, and the correlation coefficient is 0.188. This means that the more listings a host posts on Airbnb, the higher the listed price is. Price is also positively correlated with number_of_nearby_venues, which means that the more venues there are in the neighborhood, the higher the listed price is. In summary, the correlation between price and any of its predictors is either close to zero or weak.

A multiple regression model was built by using the five predictors to predict price. Because "room_type" is a categorical variable with 3 categories, 3 dummy variables were created based on that variable. The final regression model was built on price and 7 predictor variables. The model was estimated with the training data and predictions were made on the test data. Table 3 shows the summary results of the estimated model.

Table 3. Summary of Results of the Multiple Linear Regression Model



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.327
Model:                            OLS   Adj. R-squared:                  0.327
Method:                 Least Squares   F-statistic:                     1322.
Date:                Fri, 04 Oct 2019   Prob (F-statistic):               0.00
Time:                        14:40:28   Log-Likelihood:                -91939.
No. Observations:               16308   AIC:                         1.839e+05
Df Residuals:                   16301   BIC:                         1.839e+05
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         90.0604      1.096     82.159      0.000      87.912      92.209
x1            -0.1564      0.021     -7.494      0.000      -0.197      -0.116
x2            -0.0341      0.013     -2.718      0.007      -0.059      -0.010
x3             0.1798      0.011     15.897      0.000       0.158       0.202
x4             0.1124      0.012      9.154      0.000       0.088       0.137
x5            94.9697      1.084     87.624      0.000      92.845      97.094
x6             6.0977      1.106      5.514      0.000       3.930       8.265
x7           -11.0070      2.607     -4.222      0.000     -16.117      -5.897
==============================================================================
Omnibus:                     2915.610   Durbin-Watson:                   1.874
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5316.239
Skew:                           1.130   Prob(JB):                         0.00
Kurtosis:                       4.648   Cond. No.                     3.53e+17
==============================================================================
```

Collectively, the 7 predictors accounted for about 33 percent of the total variance in price. In addition, the coefficient associated with each of the predictors was significant at the 0.05 level.

The R-square was 0.34 on the test data, which indicates that the model performed fairly well.

Discussion

Explorations and visualizations of the New York City Airbnb data provide the following interesting insights:

1. The scatterplot between room type and price suggests that room type has little influence on price. In other words, price is not driven by room type.
2. There is a negative correlation, although very small, between price and number of reviews. For example, some listings that are priced at 400 do not have a single review. On the contrary, quite a few listings have more than 300 reviews but are priced below 200. A plausible explanation may be that rooms that are less expensive tend to be booked by more people, which leads to more reviews.
3. Minimum nights does not seem to have any impact on price. In other words, requiring the minimum length of stay does not influence how the room is priced.
4. The total number of listings posted by the provider is positively correlated with price. This means the more rooms a provider posted on Airbnb, the higher the room price is. A possible explanation for this finding is that providers with multiple listings are better at setting the price to their own advantage. Another explanation is that providers with more listings tend to provide better services, which drive up the price.
5. Number of nearby venues is positively correlated with price. This is not hard to understand. Location or neighborhood is the main factor that drives the price of a room (or home).

In summary, although the correlations between price and the predictors are weak at best, the predictors, as a whole, explained a relatively big proportion of the variance in price. In addition, if data for all the five boroughs were investigated or if the outliers were not excluded, the resulting R-square would be very likely smaller.

Conclusion

Obviously, price of a listing on Airbnb can be influenced by many factors other than those included in this study, such as square feet, amenities, distance to popular tourist attractions, accessibility to public transportation, and whether photos of the room (or home) are provided. Data on these factors, when obtained, can be used as predictors in future studies.