

Predictive Model on Airbnb Price in New York City

Yuze Gu^{1, a, *}

¹*Springside Chestnut Hill Academy, 500 W Willow Grove Avenue, Philadelphia, United States*
a. ygu@sch.org

**corresponding author*

Keywords: economics, Airbnb ecosystem, hotels and accommodations, business, predictive model, artificial intelligence

Abstract: This study is aimed at building a predictive model for Airbnb prices, specifically focused on New York City. To ensure the accuracy of the model and the corresponding results, public governmental data was used in this study. All the records derived from the database were then randomly assigned into 2 groups: training sample (50%) and testing sample (50%). A linear regression model was built to predict the length of stay in a Medicare skilled nurse facility in 2015 using the training sample and then was applied in the testing sample for performance assessment. The results turned out to be stunning. Areas such as Bronx, Brooklyn, Queens, Staten were found out to be significantly less expensive than Manhattan. Private rooms, Shared rooms were less expensive than the whole home/house. These predictors can then be used to build the predictive model. Multiple R-squared was 0.5179 and the adjusted R-squared was 0.5176. The average mean squared error for the linear model in the testing sample was 0.21. The correlation between the predicted and the observed was 0.72. The min-max accuracy was 0.93. The mean absolute percentage deviation was 7.22%. In general, we identified the importance of predictors of Airbnb price in New York City, such as location, type of room, number of reviews, etc. Our predictive model suggested that areas such as Bronx, Brooklyn, Queens were among the lowest priced rooms and shared rooms were generally less costly than renting a whole apartment.

1. Introduction

Airbnb Inc. is used to mostly coming out on top. Despite fierce opposition from traditional hoteliers, the home-rental startup has negotiated regulations that have allowed it to operate in more than 500 cities, from San Francisco to Tokyo. Now the company, last privately valued at \$31 billion, is gearing up to go public [1][2]. But it's hit a wall in New York City, where the rules are stricter, and Airbnb's often aggressive tactics appear to have backfired [3].

The New York City Council voted unanimously to significantly restrict Airbnb and other online home rental services, joining a growing movement of cities around the globe in stepping up regulation of the so-called sharing economy [2][4]. The bill aims to prevent landlords and tenants from illegally renting out apartments for a few days at a time to tourists, a phenomenon that the city says has aggravated the housing crisis by making short-term rentals more profitable than long-term leases. If

the bill is signed into law, New York will join cities like San Francisco; New Orleans; Barcelona, Spain; and Vancouver, British Columbia, in regulating companies that profit from facilitating short-term rentals [5].

2. Data and Methods

2.1. Data

For the purpose of this research on Airbnb prices, we found out that this dataset is publicly available for research: <https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data>. In this dataset, researchers are able to dig out more about hosts, geographical availability, necessary metrics to make, in the case of this project, a predictive model. All the records were randomly assigned into 2 groups: training sample (50%) and testing sample (50%).

2.2. Method

R-squared and linear regressions are used as ways of prediction in this study. Starting with R-squared, it is a statistical measure of how close the data are to the fitted regression line. Usually, we're familiar with the r-value, also known as the correlation coefficient, which measures the strength and direction of a linear relationship between two variables in a scatterplot. R-squared is the squared form of r-value. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. The definition of R-squared is fairly straight-forward; it is the percentage of the response variable variation that is explained by a linear model. Or:

$$R - squared = \frac{\text{Explained variation}}{\text{Total variation}} \quad (1)$$

R-squared is always between 0 and 100%: 0% indicates that the model explains none of the variability of the response data around its mean. 100% indicates that the model explains all the variability of the response data around its mean.

After calculating R-squared values, Mean squared errors (MSE) were then calculated and compared between both models. Min-Max Accuracy is defined as mean (min(actual, predicted)/max(actual, predicted)). The mean absolute percentage error (MAPE), also known as mean absolute percentage deviation (MAPD), is a measure of prediction accuracy of a forecasting method in statistics, for example in trend estimation, also used as a loss function for regression problems in machine learning.

3. Results

The random sample size is 19416 in the test sample and 19417 in the training sample, a total of 38833 records. The average ln(price) was 4.697 for the full sample, 4.702 in the training sample, and 4.692 in the testing sample.

Table 1. The synopsis of the samples taken from the database.

	min	Q1	Median	Mean	Q3	Max
Full sample	2.303	4.234	4.615	4.697	5.136	9.210
Training sample	2.303	4.234	4.654	4.702	5.136	9.210
Testing sample	2.303	4.220	4.605	4.692	5.130	9.210

NY City Airbnb Price

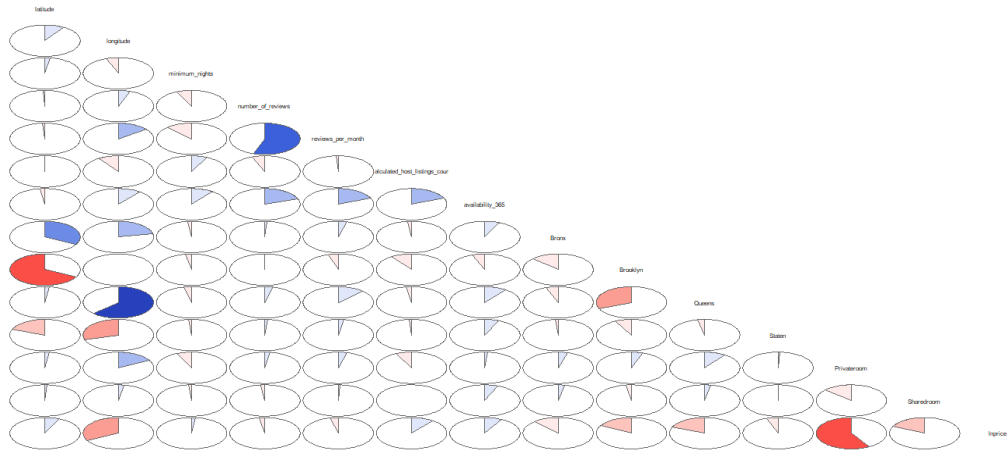


Figure 1: Pearson correlation coefficient across all the variables.

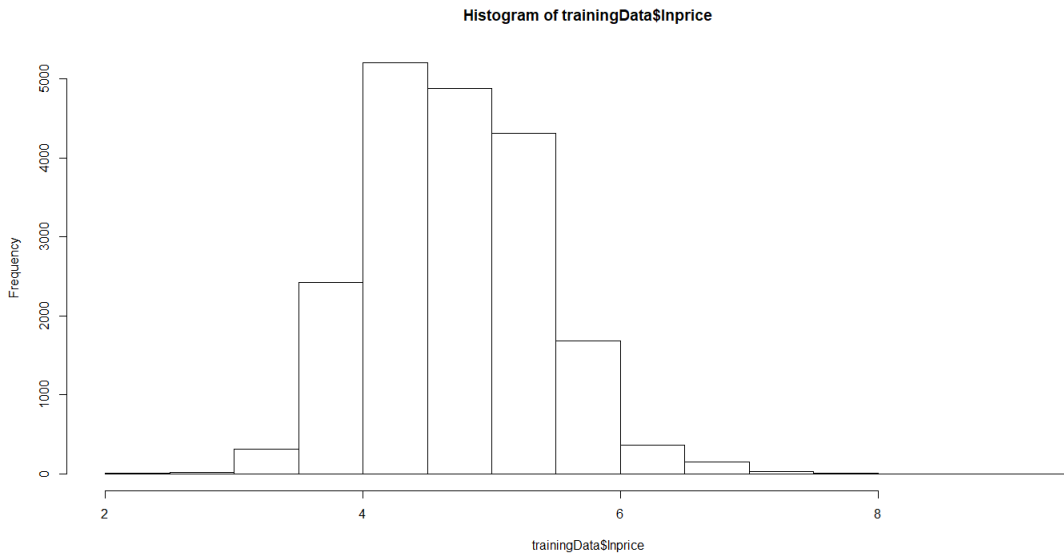


Figure 2: Distribution of The Price of Airbnb in New York City in the Training Sample.

As we can see from this histogram distribution, it is slightly unimodal, right-skewed, with possible outliers at price indicators 2 and 8. Most of the data are concentrated at price indicators 4 to 6. Now, it would be evident to also look at the distribution of the testing sample to make sure that they are of similar shapes, spread, outliers, etc. in order to ensure the accuracy of results.

Examining this histogram distribution of the test sample, we can see that the distribution is, again, unimodal, slightly right-skewed, with possible outliers near both ends. In comparison with Figure 2, the two distributions are fairly similar, which ensures that the results will be fairly accurate.

According to the linear regression, latitude, longitude, minimum nights, number of reviews, calculated host listings count, availability were significant factors for the price prediction. Amongst all areas in New York City examined, Bronx, Brooklyn, Queens, Staten were significantly less expensive than Manhattan, one of the largest economic districts in the world. One interesting insight is gained from analyzing data and the predictive model: private rooms and shared rooms were less expensive than renting the whole apartment.

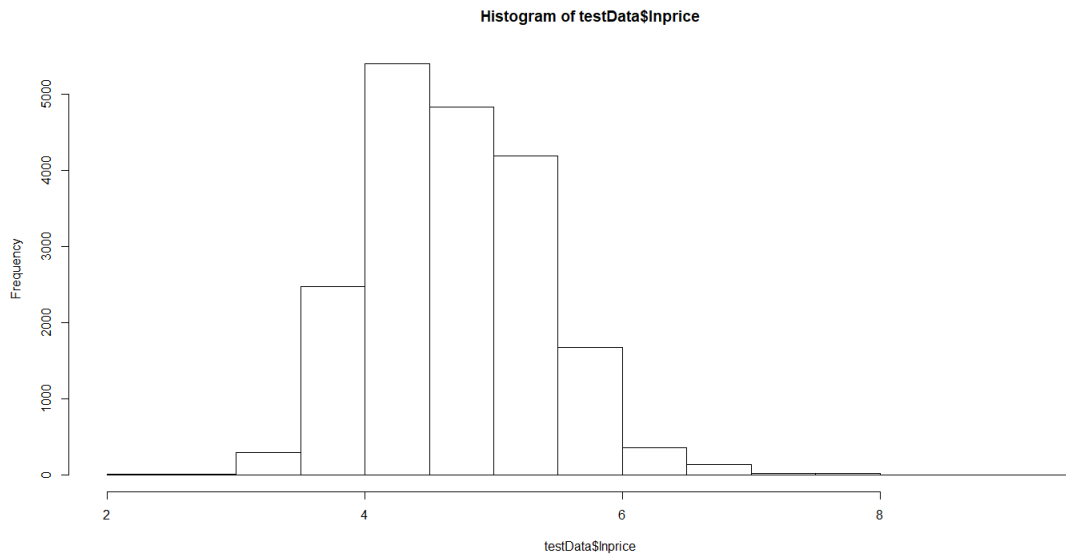


Figure 3: Distribution of The Price of Airbnb in New York City in the Test Sample.

Table 2: Linear Regression To Predict Ln(Price of Airbnb in New York City).

	Estimate	Standard Error	t value	Pr(> t)	
(Intercept)	-205.700	10.180	-20.211	< 2e-16	** *
latitude	-0.645	0.100	-6.423	0.000	** *
longitude	-3.206	0.114	-28.015	< 2e-16	** *
minimum_nights	-0.002	0.000	-10.394	< 2e-16	** *
number_of_reviews	-0.001	0.000	-6.018	0.000	** *
reviews_per_month	0.002	0.002	0.895	0.371	
calculated_host_listings_count	0.000	0.000	-2.647	0.008	**
availability_365	0.001	0.000	26.676	< 2e-16	** *
Bronx	-0.255	0.026	-9.965	< 2e-16	** *
Brooklyn	-0.287	0.012	-24.546	< 2e-16	** *
Queens	-0.165	0.017	-9.793	< 2e-16	** *
Staten	-1.152	0.042	-27.350	< 2e-16	** *
Private Room	-0.746	0.007	-108.008	< 2e-16	** *
Shared Room	-1.179	0.023	-50.300	< 2e-16	** *

Multiple R-squared was 0.5179 and the adjusted R-squared was 0.5176. The average mean squared error for the linear model in the testing sample was 0.21. The correlation between the predicted and the observed was 0.72. The min-max accuracy was 0.93. The mean absolute percentage deviation was 7.22%.

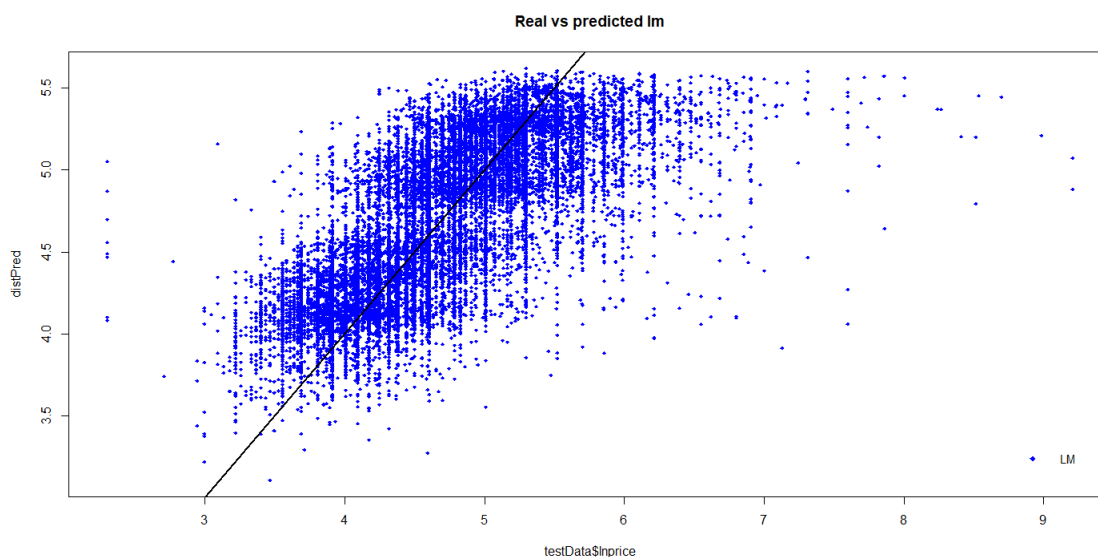


Figure 4: Predicted versus Observed Airbnb price in New York City.

4. Conclusion

According to the linear regression, latitude, longitude, minimum nights, number of reviews, calculated host listings count, availability were significant factors for the price prediction. Bronx, Brooklyn, Queens, Staten were significantly less expensive than Manhattan. Private rooms, Shared rooms were less expensive than the whole home/house.

Multiple R-squared was 0.5179 and the adjusted R-squared was 0.5176. The multiple R-squared indicated that 51.79% of the data points in the dataset can be explained by the regression line, while the adjusted R-squared suggested 51.76% of the data are explained. The average mean squared error for the linear model in the testing sample was 0.21. The correlation between the predicted and the observed was 0.72. The min-max accuracy was 0.93. The mean absolute percentage deviation was 7.22%.

In this study, we identified important predictors of various airbnb prices in New York City by sector, for location, type of room.

References

- [1] Mody, M.A., Suess, C. and Lehto, X. (2017), "The accommodation experiencescape: a comparative assessment of hotels and Airbnb", *International Journal of Contemporary Hospitality Management*, Vol. 29 No. 9, pp. 2377-2404. <https://doi.org/10.1108/IJCHM-09-2016-0501>
- [2] Baldwin Tong, Ulrich Gunter. (2020) Hedonic pricing and the sharing economy: how profile characteristics affect Airbnb accommodation prices in Barcelona, Madrid, and Seville. *Current Issues in Tourism* 0:0, pages 1-20.
- [3] Delgado-Medrano, H. M.–Lyon, K. (2016) Short Changing New York City – The impact of Airbnb on New York City’s housing market. *BJH Advisors LLC*. http://www.sharebetter.org/wp-content/uploads/2016/06/NYCHousingReport_Final.pdf

- [4] Dudás, Gábor & Vida, György & Kovalcsik, Tamás & Boros, Lajos. (2017). *A socio-economic analysis of Airbnb in New York City*. *Regional Statistics*. 7. 135-151. 10.15196/RS07108.
- [5] Javier Gutiérrez, Juan Carlos García-Palomares, Gustavo Romanillos, María Henar Salas-Olmedo. (2017). *The eruption of Airbnb in tourist cities: Comparing spatial patterns of hotels and peer-to-peer accommodation in Barcelona*. *ScienceDirect*. <https://doi.org/10.1016/j.tourman.2017.05.003>