Price Prediction and Analysis of Airbnb Data in NYC

Data 200 Empirical Project
Tingwei Li
2021/12/20

Abstract

Airbnb has been growing rapidly in recent years and it provides service for hosts and guests to share or choose short-term housing. Starting from 2008, Airbnb shows more possibilities of business models in sharing economies. Analyzing the data from Airbnb helps us better understand the phenomenon in the sharing market and provides useful suggestions for hosts and hotel managers. This paper uses the New York City dataset from website Inside Airbnb. I build models using different machine learning methods to predict price and dig out factors that influence price. During the process of predicting price, it turns out that using natural logarithm to preprocess the data can be helpful to improve the performance of the models. At the same time, the way to interpret the meaning behind statistics under the context of reality is important to the conclusion.

Introduction

Airbnb is an online platform for people to rent out their houses or apartments or even a single room to others. For guesses, they can search the listings on the Airbnb website and see the environment by the photos shared by the hosts, reviews by other guesses and other amenities previously. Thanks to the invention of the Covid-19 vaccines, the consensus to wear masks and other healthy hygiene habits, the risk of travelling has been reduced. Also, the U.S. lifted nearly 20-month international travel restrictions in November, 2021. It is predictable that travel will be the relaxation option again during the post-pandemic era. As an international student studying in Greater Boston, New York is one of the best destinations for vacation because of short distance and convenient transportation.

Until now, Airbnb has over 5.6 million active listings, 800 million people have stayed at these housings, and covering over 100 million cities. There is a giant market in Airbnb and sharing economy. It is undoubtable that Airbnb has gained great recognition as a way to travel for its features, including price advantage compared to hotels, convenience and closer to local life. If I visit New York, Airbnb would be my first housing choice.

To come up with more details about Airbnb daily housing prices in New York, I use Airbnb's New York dataset to analyze the daily price changes from various factors such as availability, neighborhoods, room types and so on. Also, I use some regression models (including linear regression, PC regression, XG Boost, Random Forest) to predict the price. Using natural logarithm to price can greatly improve the performance of the models.

Literature Review

Guttentag's survey (2018) [1] indicated that Airbnb users are attracted by the price, location and convenient house amenities. Also, the social interaction, experience are important factors for guests. Karlsson and Dolnicar (2016) [2] found that hosts were motivated by income, social interaction and sharing in Australia. Studies around the world (Chen and Xie, 2017, Kakar et al., 2018, Wang and Nicolau, 2017) [3][4][5] have tended to find that price is positively correlated with review scores, availability of the entire house, number of bedrooms, number of bathrooms, guest capacity, Superhost, length of time the host has been a member, certain amenities (e.g., parking), and number of photos. In contrast, price was typically negatively related to distance from the city's center, hosts having multiple listings, more flexible cancellation policies, instant booking availability, and the number of reviews.

Machine learning methods are widely used in predicting. Study (Zhu, Li, R., & Xie, Z., 2020) [6] shows that among the common machine learning models, including linear regression, generalized additive model, deep neural network, random forest, XGBoost and bagging (merging the previous five methods), XGBoost and random forest have a great performance in Airbnb data. During the process of feature selection, in addition to the original variables in the dataset, they create a sentiment score based on the natural language process of the name from listing. The authors highly recommended the bagging model for the topic since it can minimize the dominating power of certain specific models.

Data Description and Visualization

The dataset comes from the Inside Airbnb website and the dataset was compiled on Nov. 2nd, 2021. The summary listings dataset contains 18 variables and 37713 observations, and the detailed version listings dataset contains 74 variables and 37713 observations. After data cleaning and merging, the final dataset contains 20 variables and 37674 observations. The final variables for each observation are shown as below:

- id: the unique id for the listings
- host_id: the unique id for the host or user
- neighbourhood_group: the boroughs in NYC, including Bronx, Queens, Staten Island, Brooklyn, Manhattan
- neighbourhood: the exact neighborhood place in NYC, including
- room_type: including Entire house/apt, private room, share room, hotel room
- price
- minimum_nights: the minimum nights to stay
- reviews_per_month
- calculated_host_listings_count: the number of listings that the same host own
- availability_365: the availability in 365 days.
- number_of_reviews_ltm: the number of reviews the listing has in the last 12 months
- latitude
- longitude
- accommodates: the maximum capacity of the listing
- review_scores_rating: the rating for the listings

The histogram of price is shown in the Appendix. Figure1.1 shows a huge range of prices, but the majority of listings are from 0 to 1500 dollar. Since there are small number of extreme outliers, Figure1.1 cannot really show them. Figure1.2 processes the price with natural logarithm. The x axis represents the natural logarithm of price and the y axis is the number of listings. It is clear that Figure1.2 is a Gaussian distribution. In Figure1.2, the outliers can be clearly represented. The majority of listings lie in the range from 4 to 6 ($e^4 \approx 155$, $e^5 \approx 148$, $e^6 \approx 400$).

Map of Price (Figure 3) plots the listings under $1000. The color and circle size represent the level of the price. The detailed price distribution by boroughs is shown in the following boxplot. It shows that Manhattan has the highest average price. The Airbnb listing prices in Staten Island and Brooklyn are lower than in Manhattan but slightly higher than the prices in the Bronx and Queens.

Figure 4 and Figure 5 show the counts by boroughs and room types. Manhattan and Brooklyn account for the majority number. As for room types, Entire home/apt and private room account for more than 99% of listings in New York City Airbnb.

Model

Linear Regression

Denote Y as price, β is the coefficient vector and X represents the variables, ε is the unobserved random variable. In this multi-linear regression, $X_1 X_2, ..., X_{11}$ refers to variables of neignhourhood_group, room_type, minimum_nights, reviews_per_month, calculated_host_listings_count, availability_365, reviews_scores_rating, respectively.

$$Y = X\beta^T + \varepsilon, \beta = (\beta_0, \beta_1, ..., \beta_{11})^T, X = (1, X_1 X_2, ..., X_{11})$$

Table 2.1 shows the outcome of using linear regression by R. Since neignhourhood_group and room_type are categorical variables, R will automatically choose one category in each categorical variable separately as the baseline model. In this model, R chooses Bronx and Entire Room/apt as the baseline, so in the coefficients table, it does not show these two variables. As Table 2 shows, the R-squared is low and only reaches 0.1284. It is a very poor model for predicting the price.

Principal Components Regression

Principal Components Regression (PCR) is the regression method based on Principal Components Analysis (PCA). PCA is the method that uses orthogonal linear transformation that transforms the data to a new coordinate system but preserves the greatest variance by some scalar projection of the data. (Jolliffe, I. T., 2002) [7] Instead of using dependent variables in the Linear Regression, Principal Components Regression uses the principal components as regressors, which will be helpful especially if some dependent variables have multicollinearity problems in multi-linear regression. In this dataset, Figure 6.1 shows that the model considers 16 components to cover all the variances and the first 6 principal components explain 61.68% variance. But the R-squared is not good too, it only has 0.08 using the first 6 principal components.

Random Forest

Random Forest is an ensemble learning method and it can be used as both regression and classification problems. It is a bagging method that combines a certain number of decision trees. For regression tasks, the mean or average prediction of the individual trees is returned (Ho TK, 1998 [8]). Random Forest can effectively overcome the disadvantage of overfitting in decision tree and provides more flexibility. In the NYC Airbnb dataset, the R-squared equals 0.1953 when using Random Forest to predict the price (ntree = 50).

XG Boost (Extreme Gradient Boosting)

XG Boost comes from GBDT (Gradient Boosting Decision Tree) algorithm but using more regularized formalization to control overfitting. Also, Szilard Pafka’s(2015) [9] study indicated that XG Boost is faster when compared to other implementations of gradient boosting. This machine learning algorithm shows great success in Kaggle competition. In NYC Airbnb dataset, the R-squared equal to 0.3084 when using XG Boost to predict the price.

Improvement of Models (with logarithm process to price)

It is exciting to find that all these machine learning algorithms experience a great improvement when using logarithm of prices. In linear regression (Table 2.2), R-squared equals to 0.5339, which is a great improvement compared to without the process of logarithm. Because the price distribution is highly unbalanced, the linear model learns too many noises in the former regression and leads to less accuracy of the prediction. However, using natural logarithm makes the predict variables more continuous and similar to the normal distribution. Thus, the linear model will be more accurate.

Principal Components regression also improves greatly (Figure 6.2). The R-squared curve shows the same trend compared to the one without logarithm (Figure 6.1) but the R-squared start from the higher level. Both Random Forest and XG Boost gain improvement using logarithm. For Random Forest, R-squared increases from 0.1953 to 0.6631. For XG Boost, R-squared increases from 0.3084 to 0.6652. Table 3 shows the comparison between these four algorithms.

Empirical Analysis

In the correlation heatmap (Figure 7), availability shows a positive correlation with price. In multi-linear regression, Table 2.1 also indicates that the variable availability_365 has a positive estimated coefficient 0.122 and has a high significance level, which means that if the day of availability increase, the price will increase based on the multilinear regression formula.

$Y = X\beta^T + \varepsilon, \beta = (\beta_0, \beta_1, ..., \beta_{11})^T, X = (1, X_1 X_2, ..., X_{11})$.

Apart from comparing in a linear model, Tabel 4.1 - Table 4.6 shows the detailed correlation test and control of the variables of boroughs and room types.

According to the theory of supply and demand (Figure 8), if supply exceeds demand for a good or service, prices will fall (and vice versa). Availability can be a way to measure popularity, which can be explained as demand. If availability is low, demand is high (and vice versa). Starting from this theory, price should be higher if availability is low because demand exceeds supply. Price should be lower if availability is high because supply exceeds demand. In other word, availability should have an inverse relationship with price.

However, as the tables show, in the whole New York City area, Entire home/apartment and private rooms have a positive correlation with prices at high significant level. If breaking into boroughs, availability and price show a positive correlation in Manhattan, Bronx, Brooklyn and Queens. It would be counterintuitive that the hosts set up a high price when noticing their houses are not popular and it contradicts with the Supply-Demand theory. Generally, lower prices can attract more people to choose listings.

Considering the context of reality, the reason can be that since the high price of listings, fewer users choose these listings but they more probably choose the lower price listings. Thus, higher price listings remain high availability. In other words, price and availability still have an inverse correlation, but since the dataset only shows the result instead of change, the relationship between price and availability cannot be reflected on the calculation of correlation.

At the same time, neither in the whole New York City area nor in any borough hotel rooms and shared rooms show a high significant level of availability and price. It is believed that the availability of hotel rooms and shared rooms have little influence on the price. Due to the limited listings in Staten Island, the p value is not good enough to reveal the relationship between availability and price.

Conclusion

Data cleaning and processing is crucial to models' performance. In this New York City Airbnb dataset, using natural logarithm to price can effectively improve different regression models. This preprocessing method can apply to the data containing extreme outliers in the regression assignments. XG Boost and Random Forest have performance in price prediction. Understanding the context behind data is also important. In New York City, the price of Airbnb listings shows certain relationships that are easily misunderstood with availability. Considering reality will help us to correct the understanding of the phenomenon.

References

[1] Guttentag, Smith, S., Potwarka, L., & Havitz, M. (2018). Why Tourists Choose Airbnb: A Motivation-Based Segmentation Study. Journal of Travel Research, 57(3), 342–359.
[2] Karlsson, & Dolnicar, S. (2016). Someone’s been sleeping in my bed. Annals of Tourism Research, 58, 159–162.
[3] Chen, Yong, and Karen Xie. “Consumer Valuation of Airbnb Listings: a Hedonic Pricing Approach.” International Journal of Contemporary Hospitality Management, vol. 29, no. 9, 2017, pp. 2405–2424.
[4] Kakar, Venoo, et al. “The Visible Host: Does Race Guide Airbnb Rental Rates in San Francisco?” Journal of Housing Economics, vol. 40, 2018, pp. 25–40.
[5] Wang, Dan, and Juan L Nicolau. “Price Determinants of Sharing Economy Based Accommodation Rental: A Study of Listings from 33 Cities on Airbnb.com.” International Journal of Hospitality Management, vol. 62, 2017, pp. 120–131.
[6] Zhu, Li, R., & Xie, Z. (2020). Machine Learning Prediction of New York Airbnb Prices. 2020 Third International Conference on Artificial Intelligence for Industries (AI4I), 1–5. https://doi.org/10.1109/AI4I49448.2020.00007
[7] Jolliffe, I. T. Principal Component Analysis. 2nd ed., Springer, 2002.
[8] Ho, Tin Kam. “The Random Subspace Method for Constructing Decision Forests.” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 8, 1998, pp. 832–844.
[9] Szilard Pafka. “Benchmarking Random Forest Implementations” http://datascience.la/benchmarking-random-forest-implementations/

Appendix

Table1 Statistics Summary of Variables

Variables	Type	min	1st Qu	Median	Mean	3rd Qu	Max
id	character						
Host_id	character						
Neighbourhood_group	character						
neighbourhood	character						
Price	numerical	10.0	69.0	111.0	165.6	181.0	10000.0
Minimun_nights	Numerical	1	3	30	22.14	30	1250
Reviews_per_month	numerical	0.01	0.19	1.03	1.64	1.641	136
Calculated_host_listings_count	Numerical	1	1	1	16.24	3	391
availability_365	Numerical	0	0	67.5	129.9	295	365
number_of_reviews_ltm	Numerical	0	0	0	4.667	3	624
latitude	Numerical	40.5	40.69	40.73	40.73	40.76	40.91
longitude	Numerical	-74.25	-73.98	-73.95	-73.95	-73.93	-73.71
accommodates	Numerical	1	2	2	2.786	4	16
review_scores_rating	numerical	0	4.585	4.680	4.585	4.950	5.000

Figure 1.1

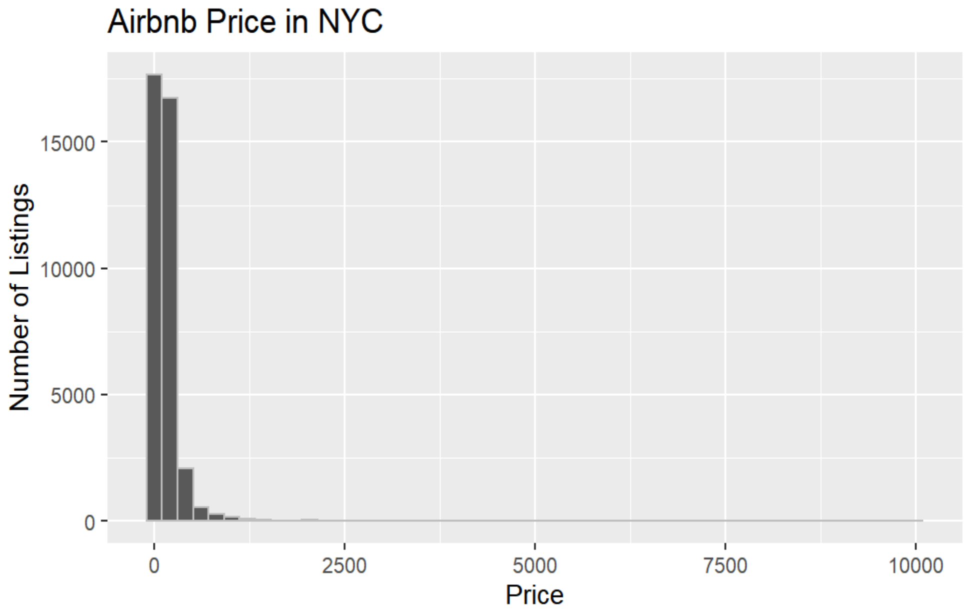


Figure 1.2

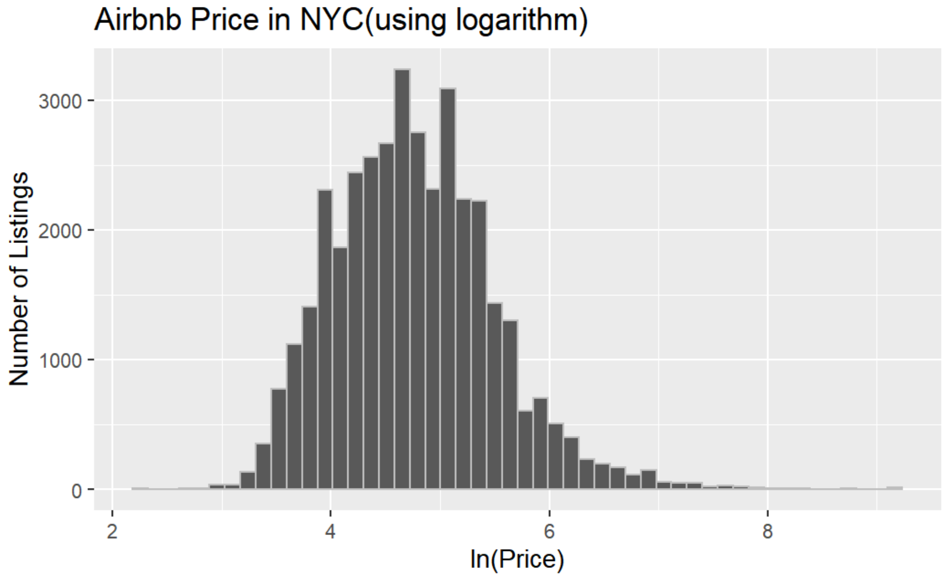


Figure 2

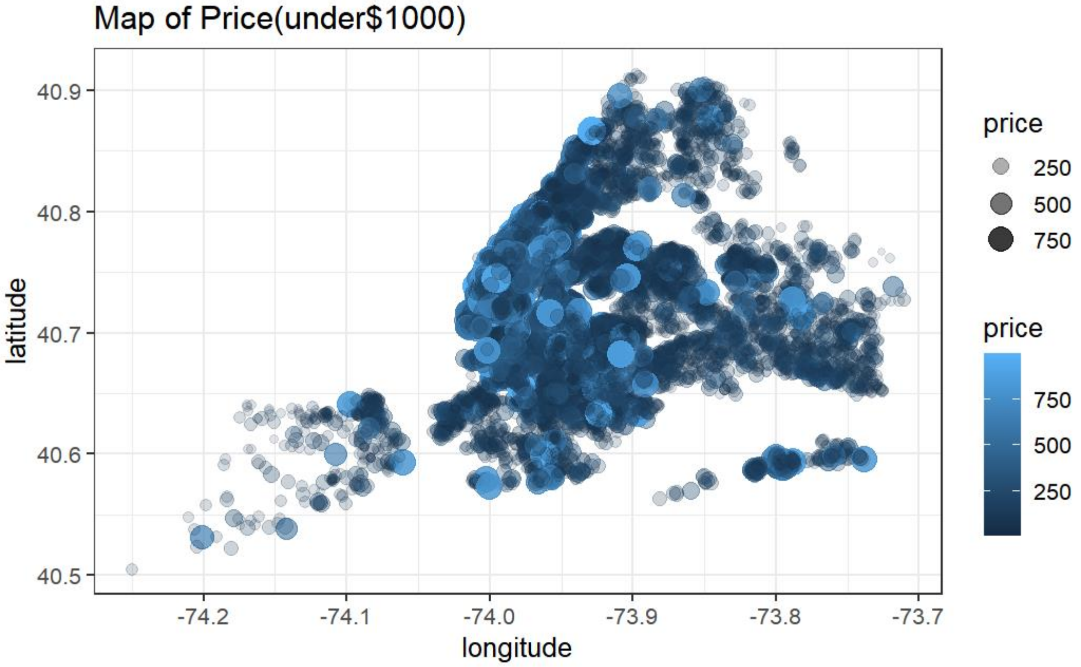

Map of Price(under$1000)
latitude
40.9
40.8
40.7
40.6
40.5
longitude
-74.2
-74.1
-74.0
-73.9
-73.8
-73.7
price
250
500
750
price
750
500
250

Figure 3

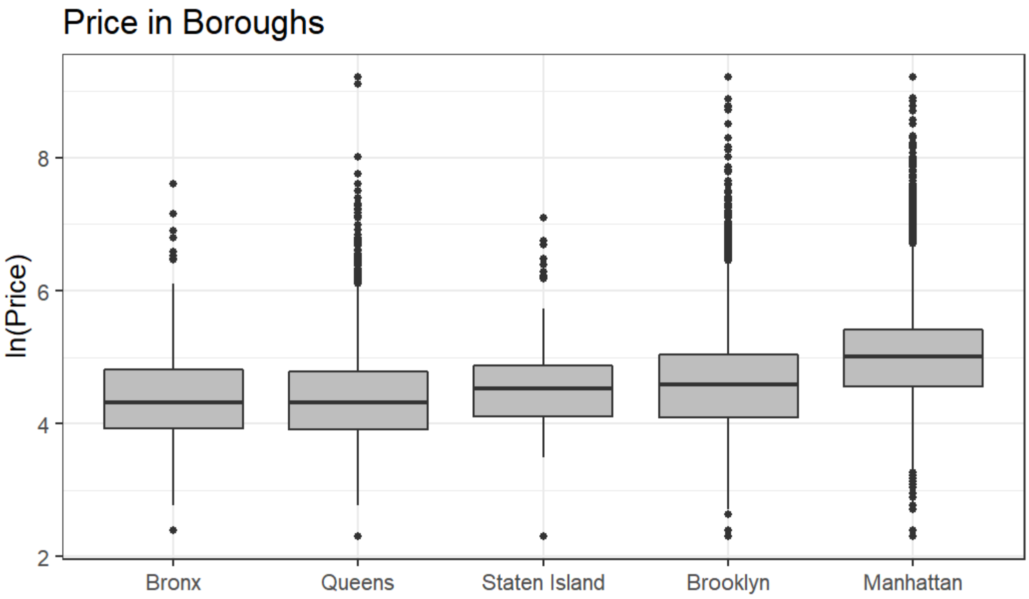

Price in Boroughs
ln(Price)
8
6
4
2
Bronx
Queens
Staten Island
Brooklyn
Manhattan

Figure 4

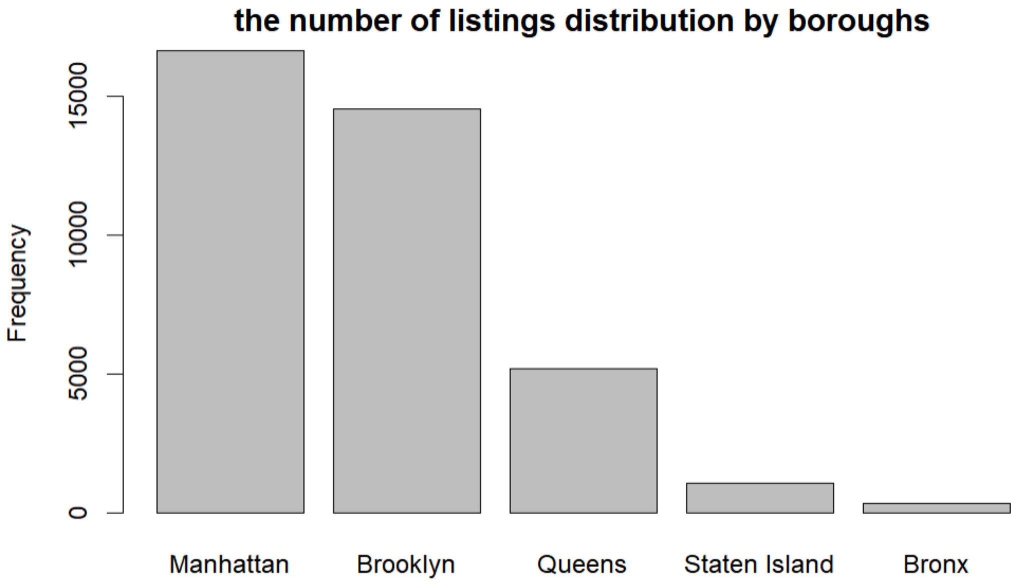

the number of listings distribution by boroughs
Frequency
0
5000
10000
15000
Manhattan
Brooklyn
Queens
Staten Island
Bronx

Figure 5

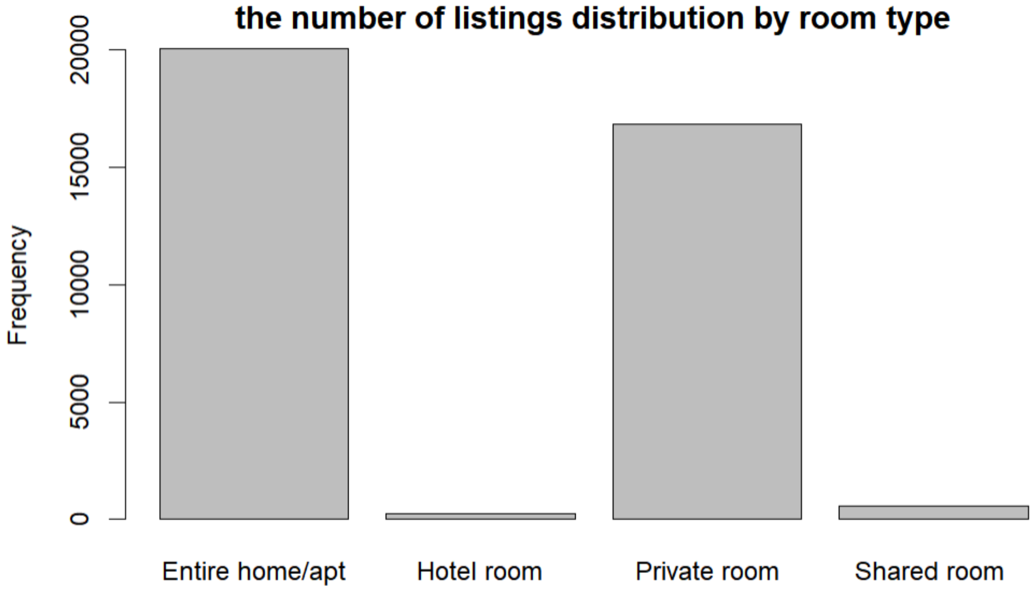
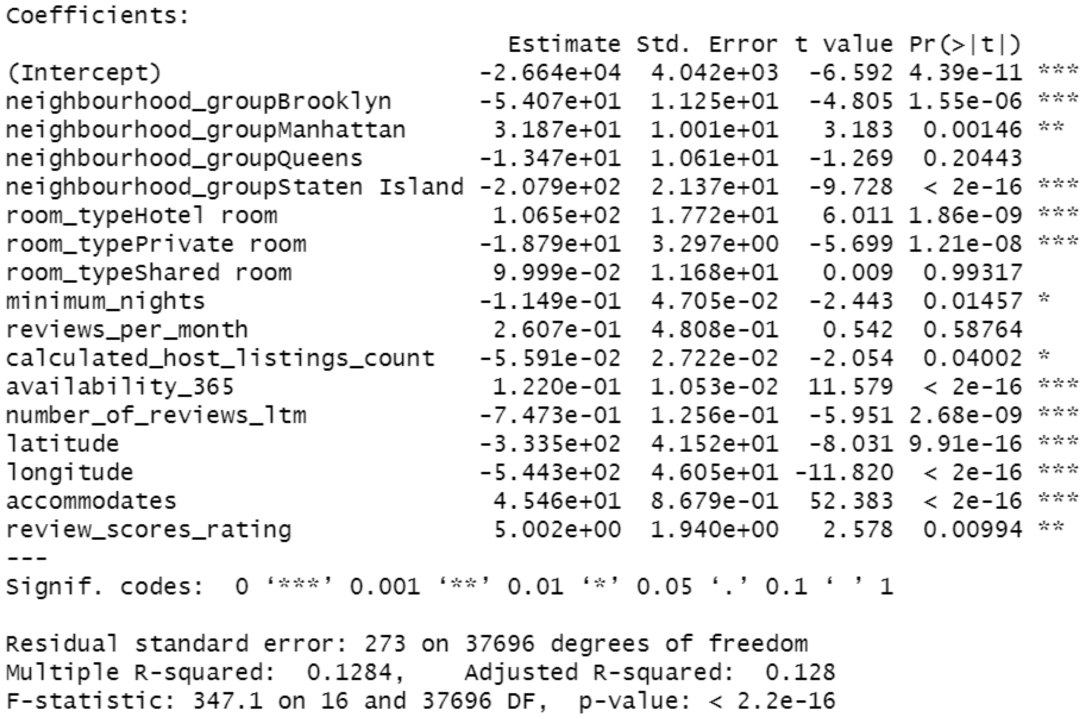

the number of listings distribution by room type
Frequency
0
5000
10000
15000
20000
Entire home/apt
Hotel room
Private room
Shared room

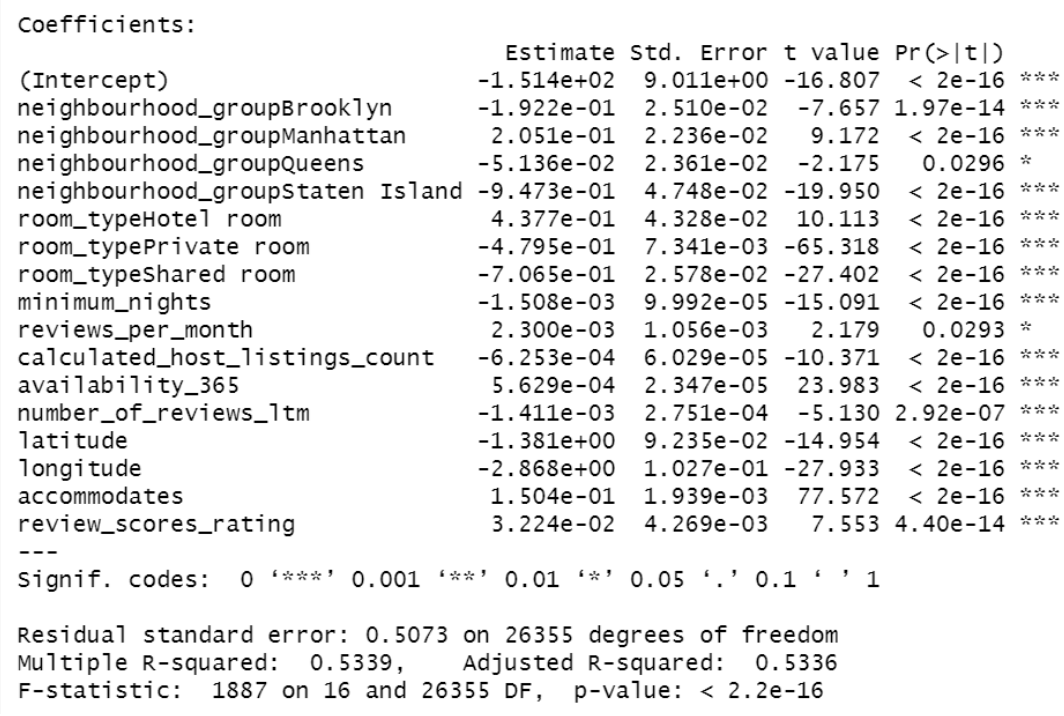
Table 2.1 Linear regression without logarithm process of price



```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     -2.664e+04  4.042e+03  -6.592 4.39e-11 ***
neighbourhood_groupBrooklyn     -5.407e+01  1.125e+01  -4.805 1.55e-06 ***
neighbourhood_groupManhattan     3.187e+01  1.001e+01   3.183  0.00146 ** 
neighbourhood_groupQueens       -1.347e+01  1.061e+01  -1.269  0.20443    
neighbourhood_groupStaten Island -2.079e+02  2.137e+01  -9.728  < 2e-16 ***
room_typeHotel room              1.065e+02  1.772e+01   6.011 1.86e-09 ***
room_typePrivate room           -1.879e+01  3.297e+00  -5.699 1.21e-08 ***
room_typeShared room             9.999e-02  1.168e+01   0.009  0.99317    
minimum_nights                  -1.149e-01  4.705e-02  -2.443  0.01457 *  
reviews_per_month                2.607e-01  4.808e-01   0.542  0.58764    
calculated_host_listings_count  -5.591e-02  2.722e-02  -2.054  0.04002 *  
availability_365                 1.220e-01  1.053e-02  11.579  < 2e-16 ***
number_of_reviews_ltm           -7.473e-01  1.256e-01  -5.951 2.68e-09 ***
latitude                        -3.335e+02  4.152e+01  -8.031 9.91e-16 ***
longitude                       -5.443e+02  4.605e+01 -11.820  < 2e-16 ***
accommodates                     4.546e+01  8.679e-01  52.383  < 2e-16 ***
review_scores_rating             5.002e+00  1.940e+00   2.578  0.00994 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 273 on 37696 degrees of freedom
Multiple R-squared:  0.1284,	Adjusted R-squared:  0.128 
F-statistic: 347.1 on 16 and 37696 DF,  p-value: < 2.2e-16
```

Table 2.2 Linear regression with logarithm process of price



```
Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      -1.514e+02  9.011e+00 -16.807  < 2e-16 ***
neighbourhood_groupBrooklyn      -1.922e-01  2.510e-02  -7.657 1.97e-14 ***
neighbourhood_groupManhattan      2.051e-01  2.236e-02   9.172  < 2e-16 ***
neighbourhood_groupQueens        -5.136e-02  2.361e-02  -2.175   0.0296 *  
neighbourhood_groupStaten Island -9.473e-01  4.748e-02 -19.950  < 2e-16 ***
room_typeHotel room               4.377e-01  4.328e-02  10.113  < 2e-16 ***
room_typePrivate room            -4.795e-01  7.341e-03 -65.318  < 2e-16 ***
room_typeShared room             -7.065e-01  2.578e-02 -27.402  < 2e-16 ***
minimum_nights                   -1.508e-03  9.992e-05 -15.091  < 2e-16 ***
reviews_per_month                 2.300e-03  1.056e-03   2.179   0.0293 *  
calculated_host_listings_count   -6.253e-04  6.029e-05 -10.371  < 2e-16 ***
availability_365                  5.629e-04  2.347e-05  23.983  < 2e-16 ***
number_of_reviews_ltm            -1.411e-03  2.751e-04  -5.130 2.92e-07 ***
latitude                         -1.381e+00  9.235e-02 -14.954  < 2e-16 ***
longitude                        -2.868e+00  1.027e-01 -27.933  < 2e-16 ***
accommodates                      1.504e-01  1.939e-03  77.572  < 2e-16 ***
review_scores_rating              3.224e-02  4.269e-03   7.553 4.40e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5073 on 26355 degrees of freedom
Multiple R-squared:  0.5339,	Adjusted R-squared:  0.5336 
F-statistic:  1887 on 16 and 26355 DF,  p-value: < 2.2e-16
```

Figure 6.1 PC Regression without logarithm process of price

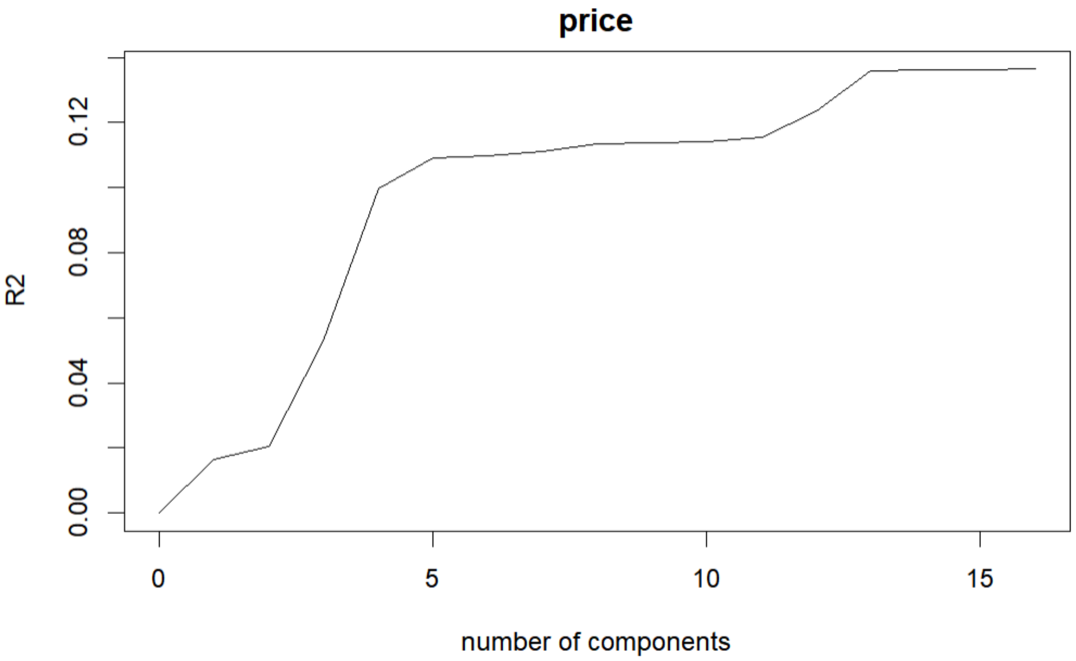


Figure 6.2 PC Regression with logarithm process of price

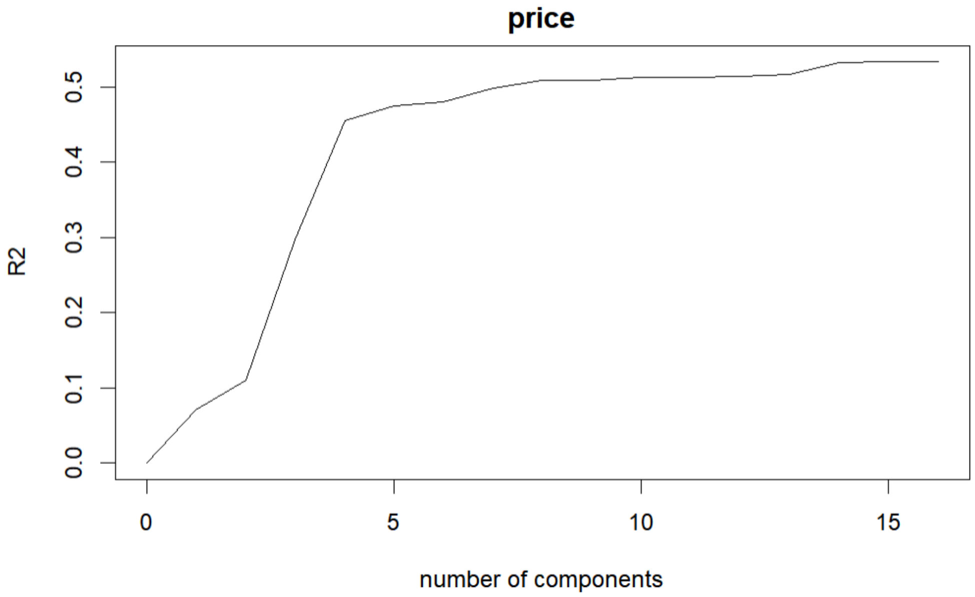


Table 3 Comparison between different models

Model	R-squared (without logarithm process)	R-squared (with logarithm process)
Linear Regression	0.1291	0.5339
PC Regression(n=6)	0.0799	0.4843
Random Forest	0.1922	0.6631
XG Boost	0.3084	0.6652

Figure 7 Correlation heatmap of variables

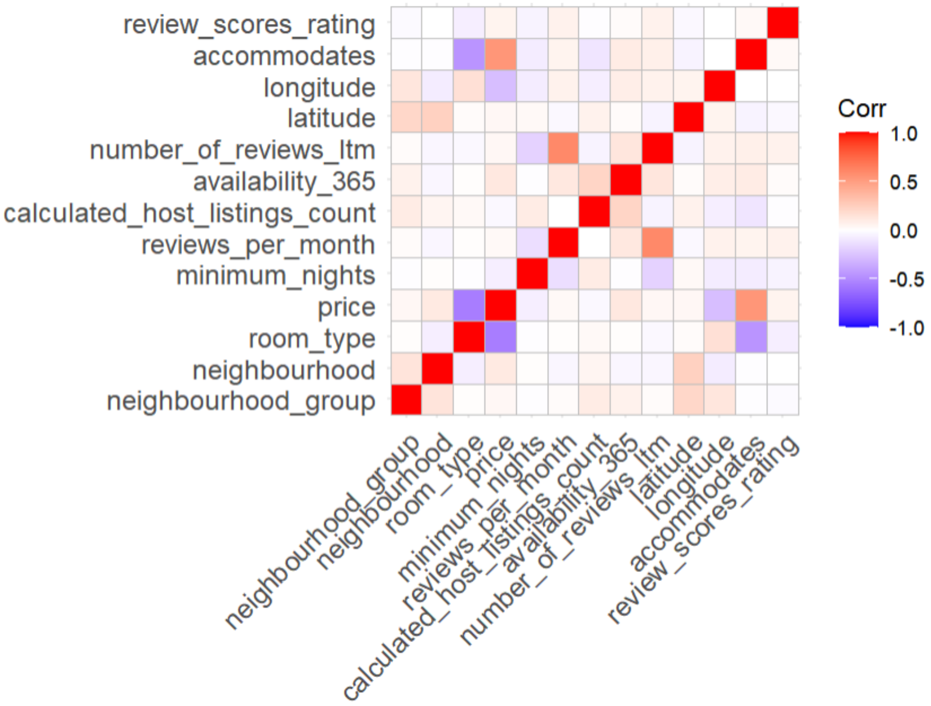


Table 4.1

All NYC area			
Room Type	Count	Correlation Coefficient	p-value
Entire home/ apt	20063	0.0886	0
Private room	16828	0.0790	0
Hotel room	207	-0.0242	0.7297
Shared room	576	0.0583	0.1808

Table 4.2

Manhattan			
Room Type	Count	Correlation Coefficient	p-value
Entire home/ apt	10062	0.0911	0
Private room	6092	0.0835	0
Hotel room	191	-0.0423	0.5611
Shared room	247	0.0475	0.4567

Table 4.3

Brooklyn			
Room Type	Count	Correlation Coefficient	p-value
Entire home/ apt	7419	0.1063	0
Private room	6894	0.0737	0
Hotel room	7	0.4074	0.3644
Shared room	187	-0.0419	0.5689

Table 4.4

Bronx			
Room Type	Count	Correlation Coefficient	p-value
Entire home/ apt	429	0.1062	0.0279
Private room	603	0.0668	0.1013
Hotel room	0	/	/
Shared room	26	0.3081	0.1256

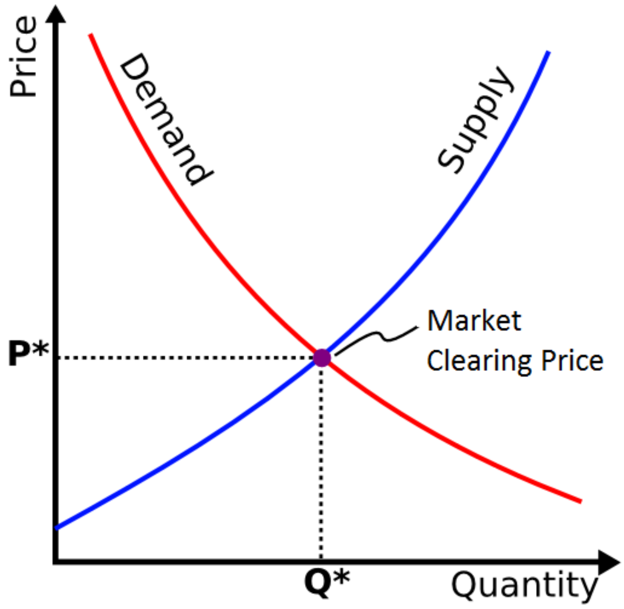
Table 4.5

Queens			
Room Type	Count	Correlation Coefficient	p-value
Entire home/ apt	1970	0.0589	0.0088
Private room	3085	0.0307	0.0873
Hotel room	9	-0.5570	0.1192
Shared room	114	0.1415	0.1332

Table 4.6

Staten Island			
Room Type	Count	Correlation Coefficient	p-value
Entire home/ apt	183	0.0456	0.5319
Private room	154	0.1326	0.1011
Hotel room	0	/	/
Shared room	2	/	/

Figure 8 Supply and Demand Curve



Code

In another file *coding.html*

Below is the pdf version printed from Chrome.