Estimating rental price for lodging, homestay or tourist houses in New York City

Capstone Project-1: Final report

By: Muluemebet G Ayalew

April, 2020

1 Introduction

The New York city Airbnb open data is chosen to be analyzed for the first capstone project. The data is provided by Airbnb. Airbnb is an online marketplace for booking and/or listing lodging, homestay or tourist houses[1].The data is accessible from the web, Inside Airbnb, and contains information about host, location, room type, price, review and availability[2]. The data contains roughly 50,000 rows and 16 columns.

The primary aim of this project is to predict the price of a new listing using the features given in the publicly available Airbnb data as well as other relevant features extracted from publicly available datasets. The results of the project benefit all three stakeholders including Airbnb, hosts and guests. Airbnb enables to recommend an appropriate price estimation for the hosts and therefore increase its revenue by increasing the number of hosts that use Airbnb's service. On the other hand, the price estimation can save hosts' time and effort to setup listing price. More importantly, if the estimated price is released publicly, hosts can attract more guests by adding some discount and gain more money. Guests can also make informed decisions when estimation is public.

The approach to address this problem is outlined in the next sections.

2 Research Questions

- The objective of the project is to make a robust model that can predict the price of the rent. To achieve this, the following research questions was addressed:
- How are the predictor variables related to price?

[1] *2020, "What is Airbnb and how does it work?"* Airbnb. Accessed Jan 30, 2020. https://www.airbnb.com/help/article/2503/what-is-airbnb-and-how-does-it-work

[2] Dec, 2019, "*Get The Data*", Inside Airbnb, Accessed Jan 30, 2020. http://insideairbnb.com/get-the-data.html

- Are the predictor variables correlated?
- Which variables are relevant to predict the price?
- Are there any other variables other than in the dataset that can affect listing price? What are they? Where can they be found?
- Which method is better for price prediction for the dataset at hand and why?
- Does the model perform well for the test data? What metrics used to measure the performance of the model

3 Methodology

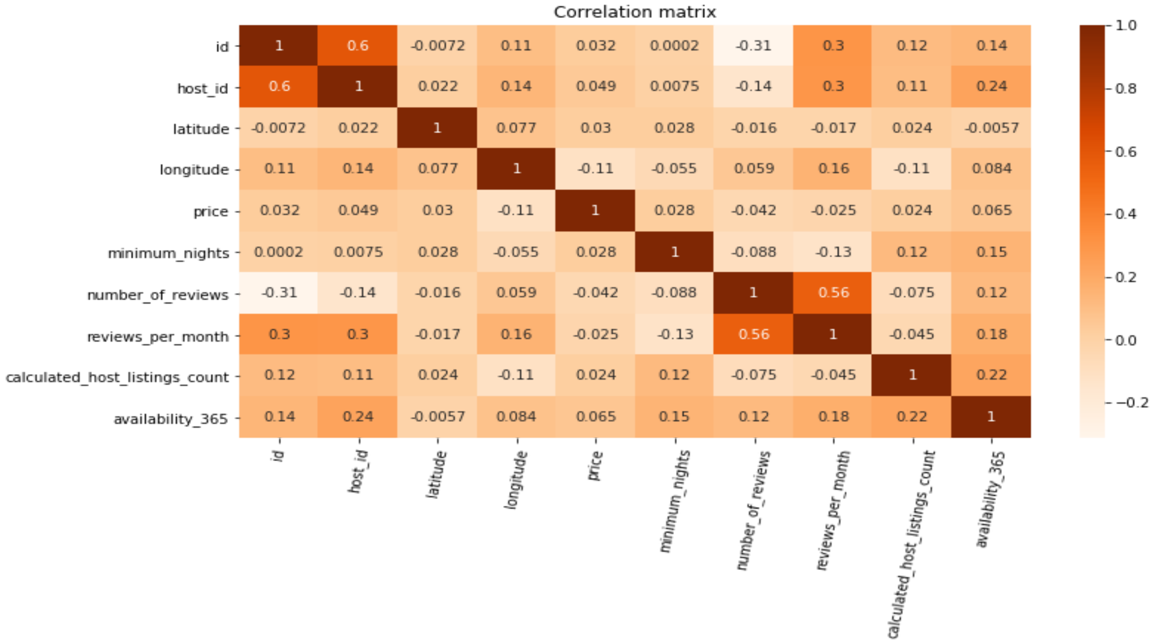
The data was explored to observe relationships between variables, outliers, missing values and summary statistics. Then, data cleaning, feature engineering, transformation or feature scaling was performed as needed. To increase the performance of the prediction, information from publicly available datasets was also included. For example, distance from the nearest subway and distance from main attractions were computed and added to the original dataset.

Once the data was prepared, regression methods such as linear regression, lasso regression, k-nearest neighbor, decision tree and random forest were used to fit the training dataset and then to test for the test dataset, split during data preparation. Then, a method better predicts the listing price was selected as the final model. The most commonly used python libraries such as pandas, numpy, matplotlib, seaborn, scikit-learn, statsmodels were used to explore, prepare, model and test the data. At the end, the code, visual stories and report is delivered.

4 Data Collection and Feature Engineering

The main data about rental listing is downloaded from the web, Inside Airbnb[3], and contains information about host, location, room type, price, review and availability. The objective is to predict the price of rental from relevant predictor variables. However, the features available from this dataset are not highly correlated with price (see Figure 1). Therefore, other datasets have been looked at and thorough feature engineering was done.

[3] Dec, 2019, "*Get the Data*", Inside Airbnb, Accessed Jan 30, 2020. http://insideairbnb.com/get-the-data.html



	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
id	1	0.6	-0.0072	0.11	0.032	0.0002	-0.31	0.3	0.12	0.14
host_id	0.6	1	0.022	0.14	0.049	0.0075	-0.14	0.3	0.11	0.24
latitude	-0.0072	0.022	1	0.077	0.03	0.028	-0.016	-0.017	0.024	-0.0057
longitude	0.11	0.14	0.077	1	-0.11	-0.055	0.059	0.16	-0.11	0.084
price	0.032	0.049	0.03	-0.11	1	0.028	-0.042	-0.025	0.024	0.065
minimum_nights	0.0002	0.0075	0.028	-0.055	0.028	1	-0.088	-0.13	0.12	0.15
number_of_reviews	-0.31	-0.14	-0.016	0.059	-0.042	-0.088	1	0.56	-0.075	0.12
reviews_per_month	0.3	0.3	-0.017	0.16	-0.025	-0.13	0.56	1	-0.045	0.18
calculated_host_listings_count	0.12	0.11	0.024	-0.11	0.024	0.12	-0.075	-0.045	1	0.22
availability_365	0.14	0.24	-0.0057	0.084	0.065	0.15	0.12	0.18	0.22	1

FIGURE 1CORRELATION MATRIX

For example, subway entrance location data, geographic location of some point of interest such as times square, Penn Station and similar data was collected from the web. Then, the distance from each listing to selected point of interest and to the nearest subway were computed and added as new predictor features.

Other feature engineering tasks such as feature encoding for categorical variables and new labeling were done. The data contains information about room type, whether it is the entire home/apartment, private room, shared room or hotel room. However, there are hotel listings that are not specified as hotel rooms. In some cases, they are listed as private rooms and in other cases entire homes/apartments. A new variable is created to distinguish a listing as hotel or non-hotel. To do this, hosting names containing the word hotel and listings listed as hotel rooms are labeled as hotel (encoded as 1) and the rest as none hotel (encoded as 0). So that we can have deeper granularity.

But this data has to be verified from Airbnb if they have detailed information about the host.

5 Data Cleaning

The listing data contains some missing values (see Table 1). There are four variables with missing values. Among these, host name and list name were not considered relevant for price prediction and were ignored. Missing values for the variable reviews per month was filled by zero. Missingness for this variable was due to the fact that the number of reviews for those listings was in the first place zero.

TABLE 1 NUMBER OF MISSING VALUES BY VARIABLE

Variable Name	Number of missing values
name	17

host_name	563
last_review	10220
reviews_per_month	10220

The price and other numerical variables such as “minimum nights” are highly skewed. The price also has zero value. One reason could be the listings were inactive and intentionally left as zero price. There could also be other reasons. In any case, including the zero price in the model might be misleading, therefore only the non- zero priced listings were considered for the analysis. Outliers are also handed by taking only data points within three standard deviations from the mean. For both training and test data the price above 1202 is considered outlier. This number is obtained by computing mean plus three standard deviations of the whole data set. A price above this value is treated as an outlier.

After all the collection, data cleaning and feature engineering tasks were done, the cleaned data was saved as a csv file for the next step of the analysis. The cleaned data was also randomly split for training (70%) and test (30%) data and saved as separate files.

6 Exploratory Data Analysis

6.1 Price distribution

The listing price is our target variable we want to predict using the model. It would be helpful to see how its distribution looks especially when linear regression is used for modeling. The following plots show price distribution of the original data, the distribution after removing outliers and log transformation in order(see from Figure 2 to Figure 6).

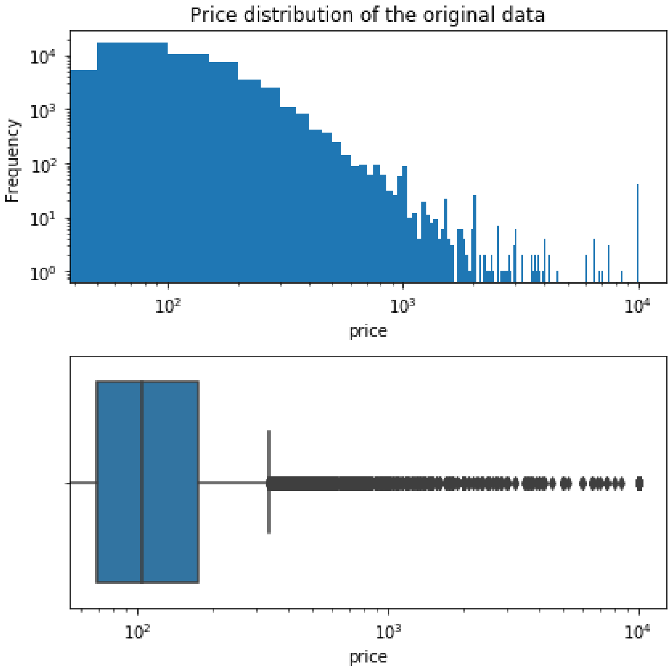


FIGURE 2 PRICE DISTRIBUTION OF ORIGINAL DATA

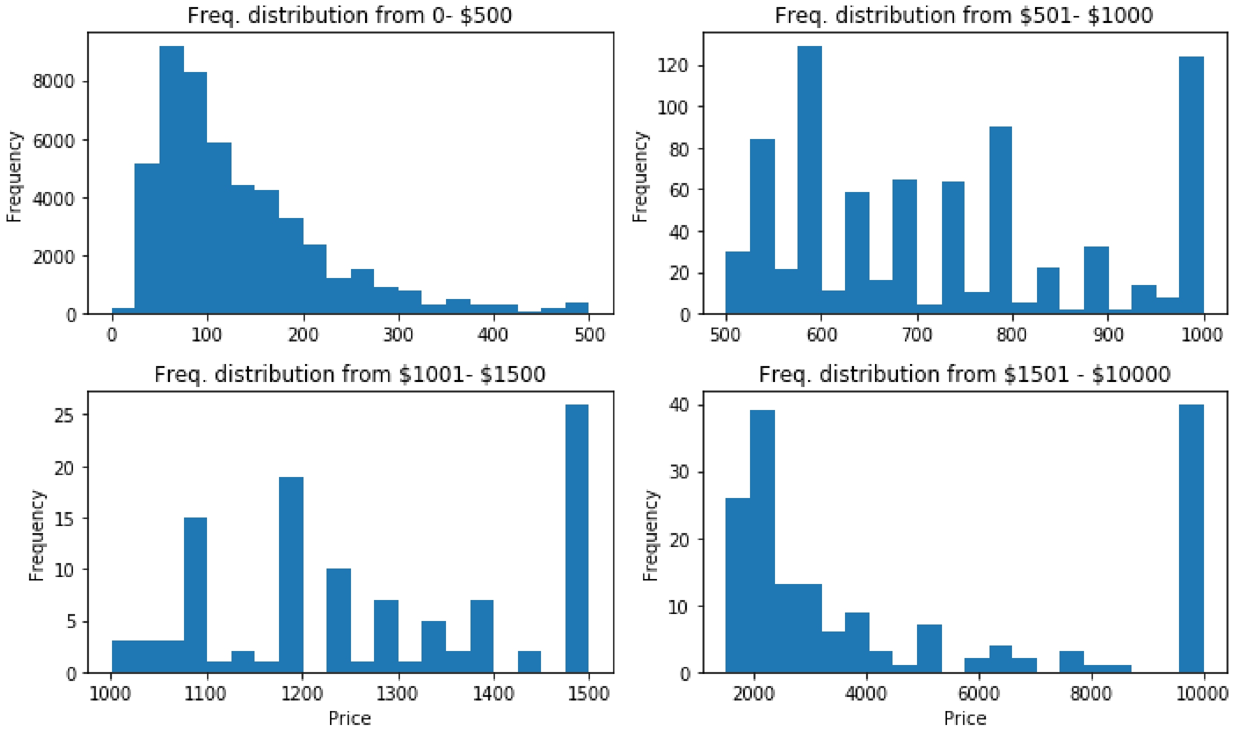


FIGURE 3 PRICE DISTRIBUTION WITH DIFFERENT PRICE RANGE

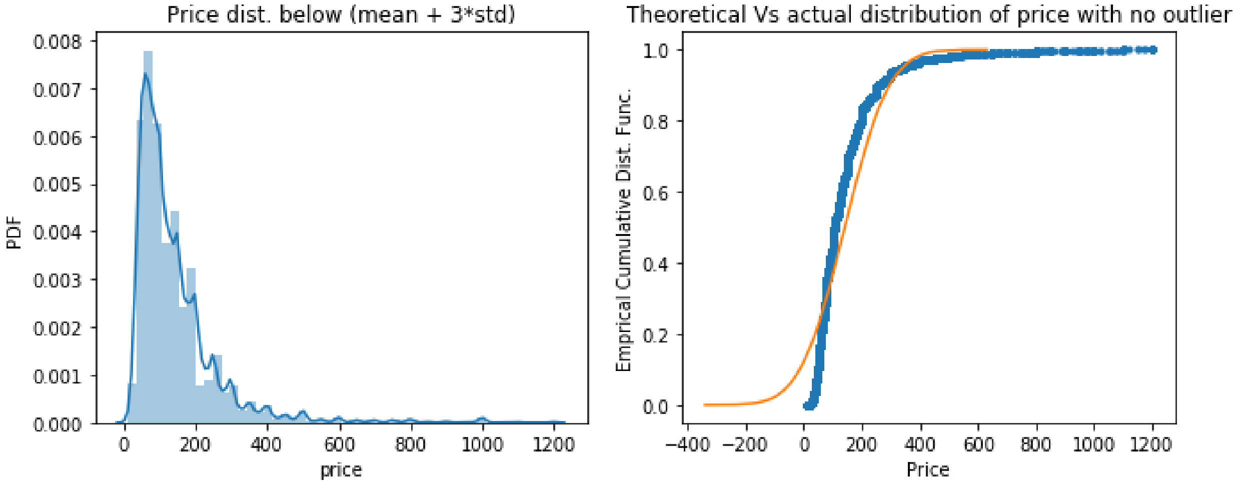


FIGURE 4 PRICE DISTRIBUTION WITHOUT THE OUTLIER

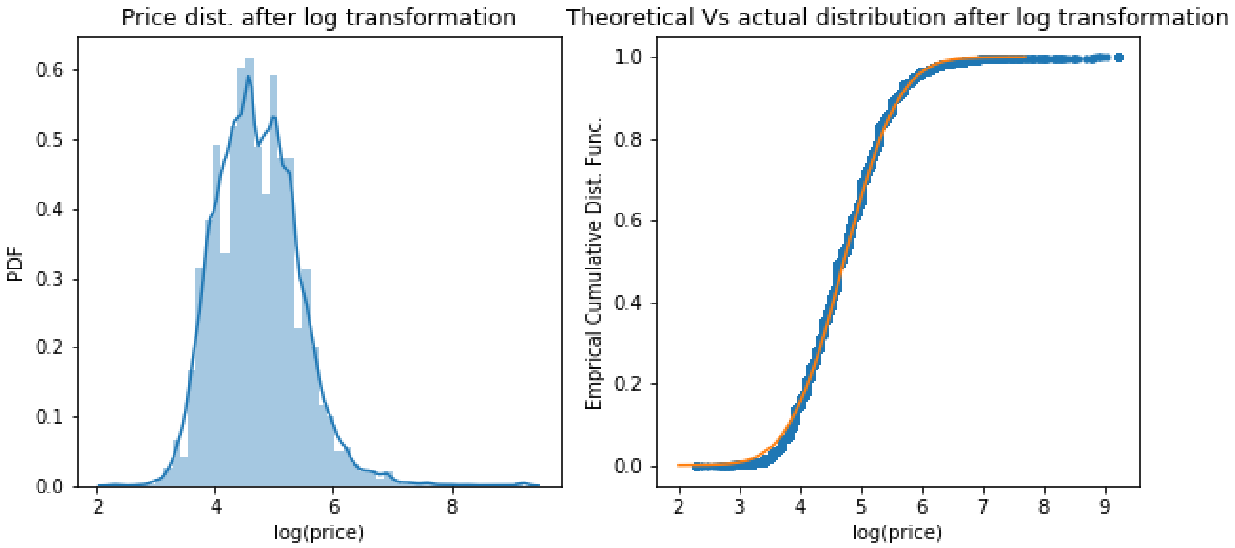


FIGURE 5 LOG TRANSFORMED PRICE

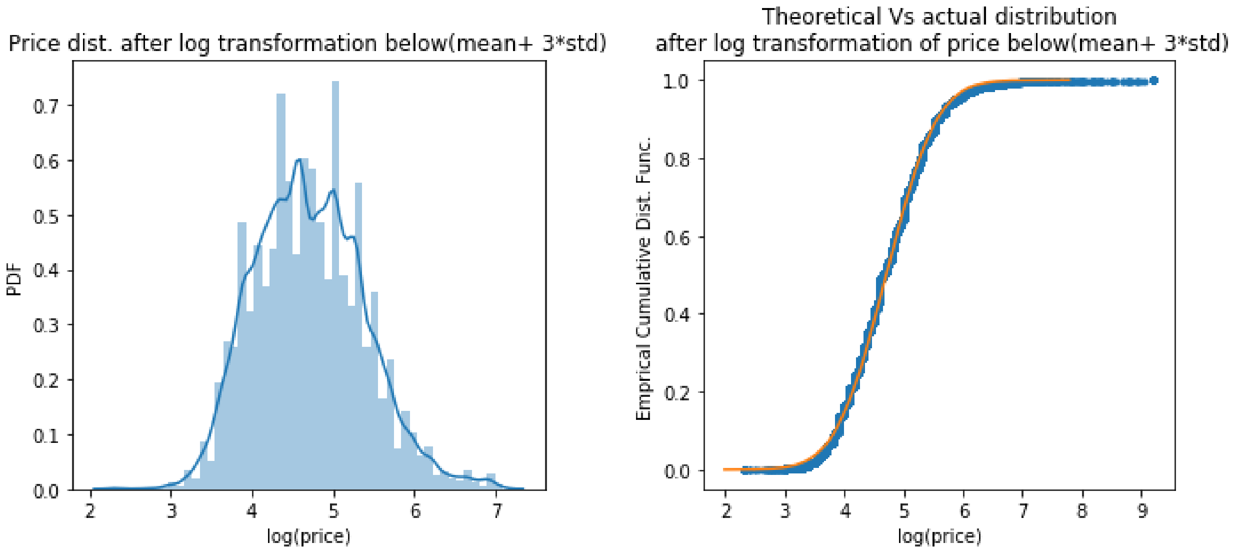


FIGURE 6 DISTRIBUTION OF LOG TRANSFORMED PRICE BELOW (MEAN+ 3*STD)

6.2 Exploration of Predictor Variables

A detailed exploration of each predictor variable was done to understand the data better and available in the notebook. However, in this report exploratory analysis for only selected features is presented.

Room Type: There are four room types; entire home/apt, private room, shared room, and hotel room. Among them, an entire home apartment is more expensive than a hotel room and private room. Shared room is the least expensive of all room types (see Figure 7).

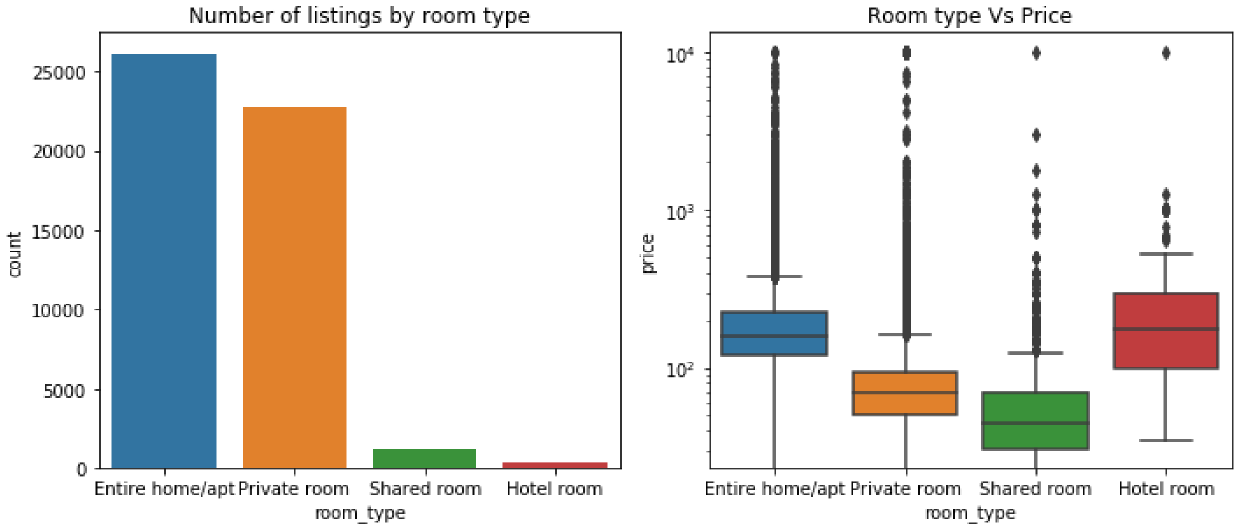


FIGURE 7 EXPLORING ROOM TYPE

Neighborhood Group: Manhattan and Brooklyn have the most listings of all and together they account more than 84 % of the listing in New York City. When observing the entire dataset, Manhattan is the most expensive, Brooklyn is the second most expensive and Bronx is the least expensive of all (see Figure 8). There are rooms with zero price in Manhattan and Brooklyn. As we can also see in Figure 9, there is no information about hotel renting in Bronx and Staten Island and any type of room is expensive in Manhattan.

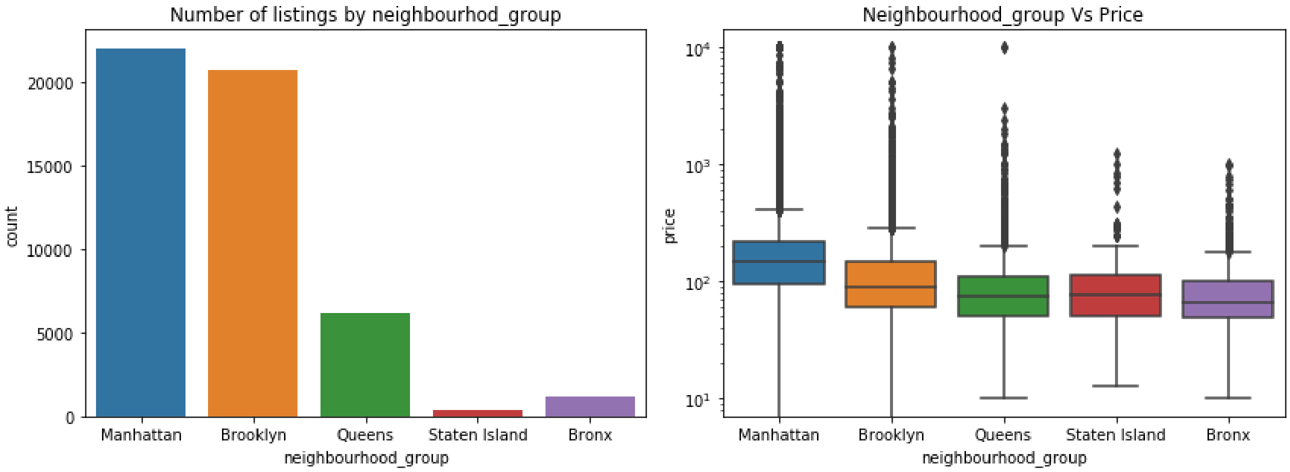


FIGURE 8 EXPLORING NEIGHBORHOOD GROUP

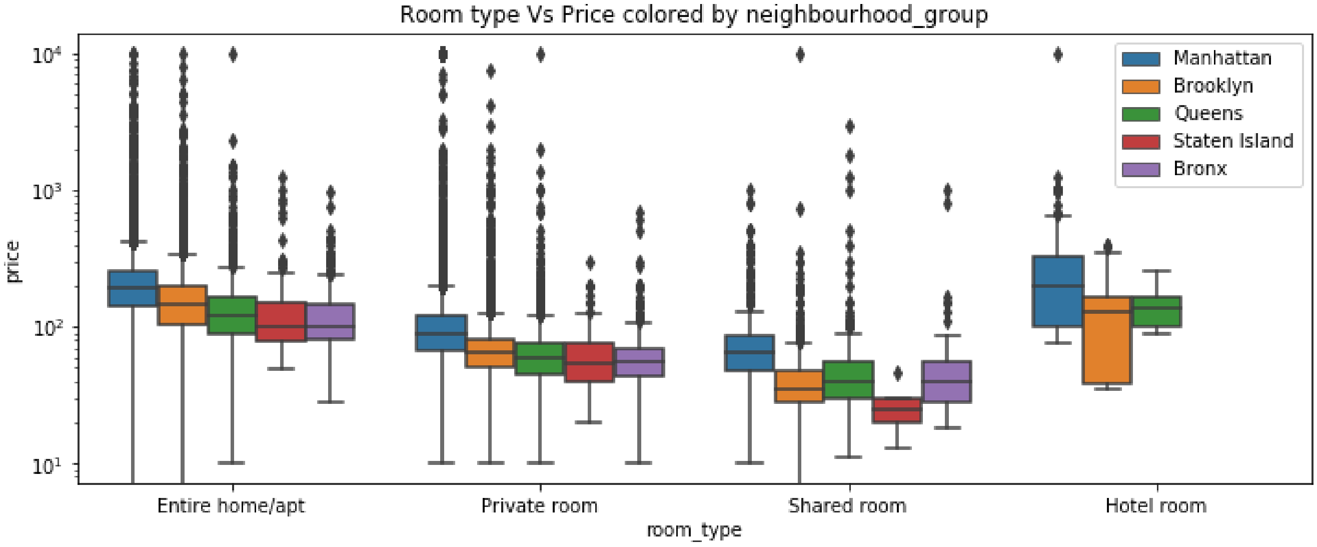


FIGURE 9 ROOM TYPE AND NEIGHBORHOOD GROUP

***Minimum Nights*:** The variable Minimum night is very skewed and multimodal (see Figure 10).

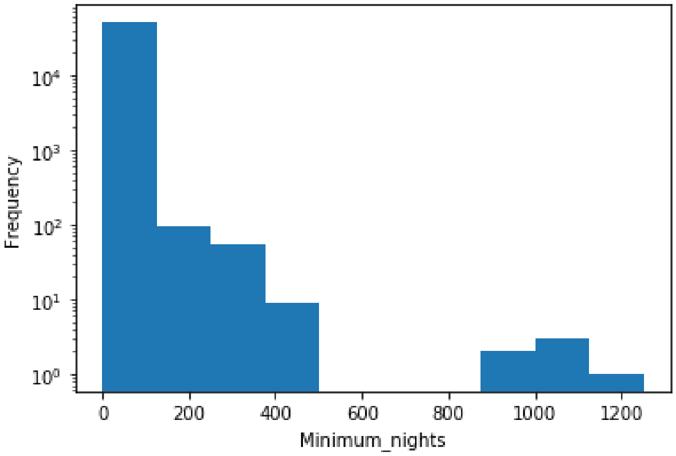


FIGURE 10 DISTRIBUTION OF MINIMUM NIGHTS

Number of Reviews （see Figure 11)

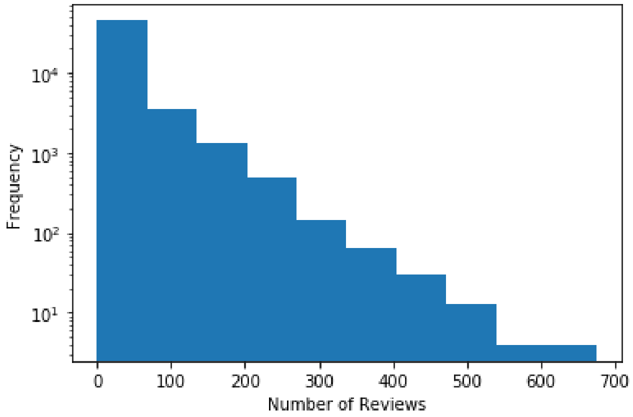


FIGURE 11 DISTRIBUTION OF NUMBER OF REVIEW

Availability 365 days: *(see* Figure 12)

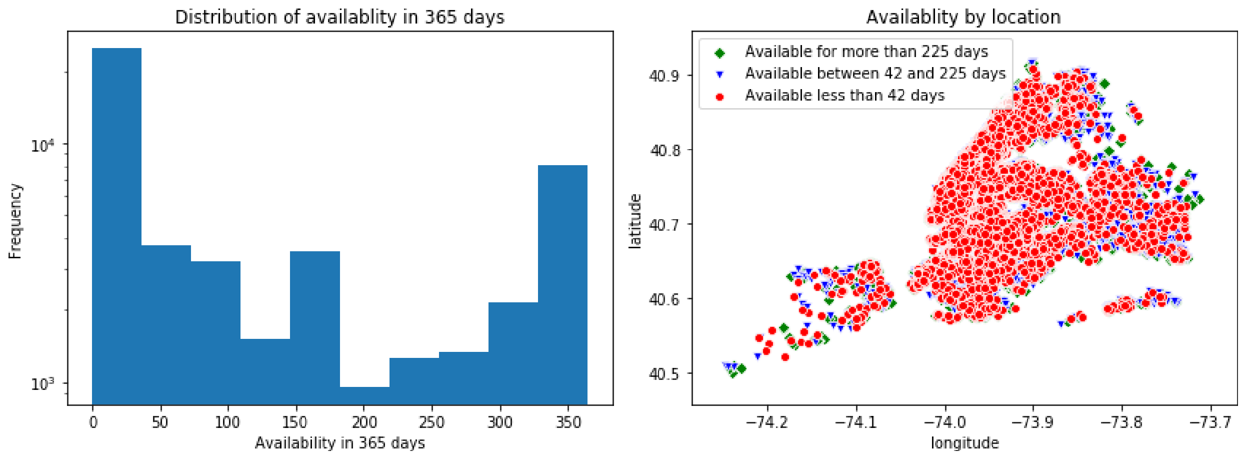


FIGURE 12 AVAILABILITY OF LISTINGS IN 365 DAYS

7 Statistical Test

Statistical test is performed to see if hotel listing price is different from non-hotel listing price. Both two -samples t-test and bootstrapped tests were performed to test the hypothesis defined as:

> The null hypothesis: The mean price of hotel and non-hotel listings are the same in New York city.
>
> Alternative hypothesis: The mean price of hotel is greater than the mean price of non-hotel listing in New York city(one-tailed)

Before performing the test, it is good to see how the distribution of the two samples look (see Figure 7).

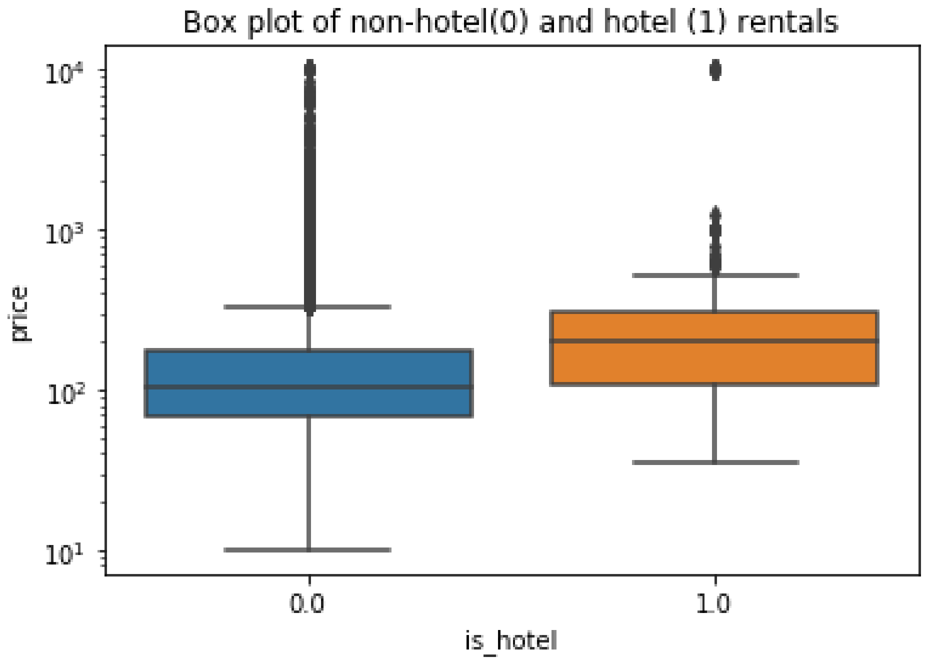


FIGURE 13 HOTEL AND NON-HOTEL LISTING PRICE DISTRIBUTION

Once the hypothesis was defined, the test statistics was chosen (t-test statistics with the same variance) and significance level was set (0.05). The assumption of same variance had to be tested before using the test statistics. Therefore, another test statistic for difference in standard deviation was tested for the two groups (hotel and non-hotel listings). Bootstrapped test was performed to test the hypothesis:

> Null hypothesis: The standard deviations of listing price for hotel and non-hotel are the same.
>
> Alternative Hypothesis: The standard deviation of listing price for hotel and non-hotel is the same (two tailed).

The p-value was 0.4736 which is greater than the significance level (0.05). The p-value is the probability to get an outcome at least as extreme as what was observed. Therefore, the difference of standard deviation of price for hotel and non-hotel listing is statistically not significant. We can accept the null hypothesis and assume that the two groups have the same variance.

After testing equal variance between the two groups, the test statistics and p-value was computed to test the mean difference (the first hypothesis). The p-value was very low for both t-test (p-value close to zero) and bootstrapped tests(p=0.0001). Concluding that the mean price of hotel listing is statistically different from the mean price of other listing. As an evidence, the bootstrapped replicates of mean difference between hotel and other listings, and the observed difference was plot (see Figure 6). The observed difference is far from other data. And thus, the p- value is very small.

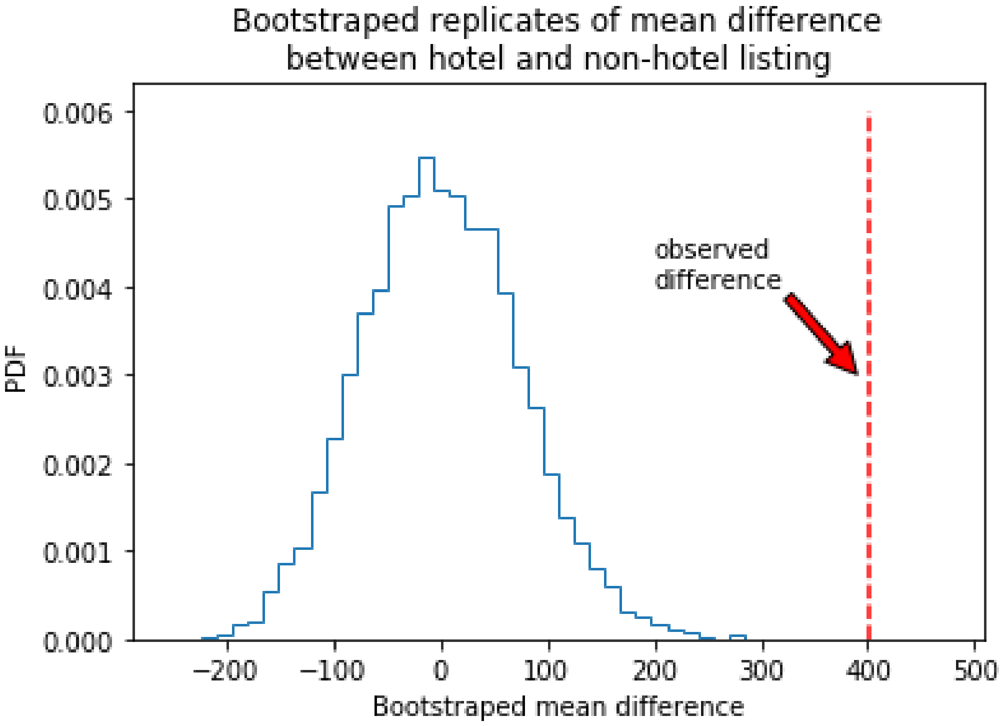


FIGURE 14 BOOTSTRAPPED REPLICATES OF MEAN DIFFERENCE

8 Modelling

Price prediction was performed using several techniques such as linear regression, lasso regression, K-nearest neighbors (Knn), decision tree, random forest and cat-boost. Depending on the type of method used, the data was prepared differently. For example, for linear regression, dummy variables are created for the variables neighborhood group and room type. And then target encoding was performed for the other categorical variable neighborhood. Since it has large categories (223), creating dummy makes the model complex. For this reason, target encoding was used. The encoding was performed after splitting the data into training and test dataset to avoid contamination. The median prices of neighborhoods were used to encode the variable. For knn model, the data was standardized to avoid scaling issue while computing distance. For decision tree and random forest models, the dummy variables created for linear models are no more appropriate for tree-based models. Thus, a different encoding was performed. For catboost model, no encoding was needed as it can accept categorical variables without encoding.

A set of variables among all the original and crafted variables were used for modeling. As already noted on the data cleaning section (section 5) of this report, some of the variable were excluded from the model because of their less importance. Variables such as distance form Time Square and distance from Penn Station are highly correlated. Only one variable, distance from Penn Station, was included in the analysis. Similarly, distance from subway entrance and distance from subway station are also correlated. Only distance from subway entrance is included in the model.

Finally, the variables neighborhood, minimum nights, number of reviews, reviews per month, calculated host listing count, availability 365, is hotel, distance from nearest subway entrance, distance from Penn Station, neighborhood groups, and room type were used for modeling.

The performance of linear regression model for test data was very low(r2=0.326). After diagnosing the assumption for linear regression and making corrective measures such as log transformation of price (outcome variable), removal of outliers and leveraging points and removal of non-significant variables (with p-value above 0.05), the model performed better for test data set(r2=0.554) (see Figure 15).

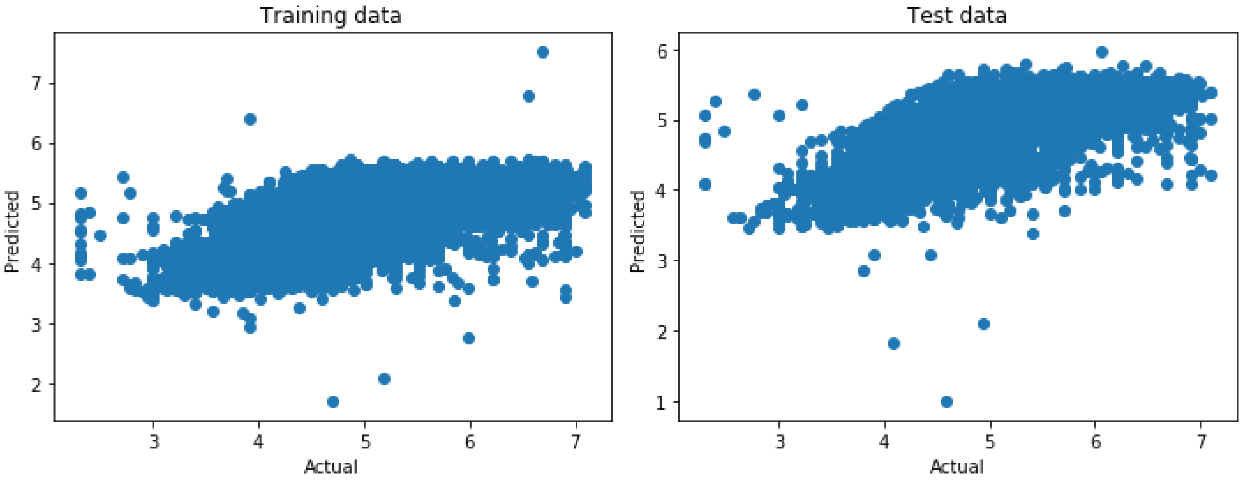


FIGURE 15 ACTUAL VS PREDICTED PLOT FOR TRAINING AND TEST DATA

Table 2 shows the summary of different models and their performance for training and test dataset. Please note that mean squared error (labeled as mean_sqr_error on the table) and mean absolute error (labeled as mean_abs_error) for knn and linear_corrected models are quite different from other models. The reason is that knn used standardized data and the error is also computed from the standardized data. The model linear_corrected is the linear model with corrective measures including log transformation of price. Since the error was computed from log transformed price, its value is also small.

TABLE 2 PERFORMANCE OF DIFFERENT MODELS

	Performance of training data			Performance of test data		
Model	**r2**	**mean_abs_error**	**mean_sqr_error**	**r2**	**mean_abs_error**	**mean_sqr_error**
linear_regresion	0.326613	55.147867	9671.908374	0.326493	56.108488	9970.002326
Linear_corrected	0.609827	0.296973	0.142955	0.553961	0.330210	0.203336
lasso	0.326527	55.090724	9673.138152	0.326217	56.041298	9974.095863
Knn	0.529588	0.379678	0.470412	0.336829	0.453797	0.663171
Decision_tree	0.383717	52.235339	8851.716164	0.351714	54.039410	9596.650074

Random_forest	0.550853	45.740185	6451.133960	0.414639	51.330724	8665.167642
Catboost	0.448902	50.097362	7915.463695	0.411079	52.008981	8717.860556

9 Summary

The linear regression model after corrective measures (log transformation of price, removal of leveraging points and outliers, and removal of non- significant variables) perform better than other models. Given the nature of the data, the supervised machine learning listed above was not performing well. Further analysis can be done by combining unsupervised and supervised learning methods. For example, cluster the data and then use cluster labels as a predictor variable and make regression. Or make regression for each cluster group. The other approach could be to create group for variables having bimodal and skewed distribution like minimum nights, availability in 365 days and price.