Analysis of New York City Airbnb and Price Prediction

Final Report

Halis Yigin

Table of Contents

1. Introduction

Airbnb is an online marketplace for arranging or offering lodging, primarily homestays, or tourism experiences since 2008. NYC is the most populous city in the United States and also one of the most popular tourism and business places in the world.

Most people who want to give their home for airbnb wonder some questions. How can I get more customers, is the location important, how should I price, how can I increase my rate? Also, customers have similar questions like hosts. They wonder if the price is good, which location they should choose for better homes . To answer these questions, I will analyze and make price predictions by using the NYC airbnb open dataset.

The project outcomes are used not only by individuals but also by companies. Data that can be analyzed and used for security, business decisions, understanding of customers' and providers' (hosts) behavior and performance on the platform, guiding marketing initiatives, implementation of innovative additional services and much more.

2. Dataset

The dataset used for this project is the https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. This dataset describes the listing activity and metrics in NYC, NY for 2019.

This data file includes all needed information to find out more about hosts, geographical availability, necessary metrics to make predictions and draw conclusions.

This dataset has around 48,895 observations in it with 16 columns. Each row represents each host. It is a mix between categorical and numeric values. The column headings are:

A. Id - listing ID
B. name - name of the listing
C. Host_id - host ID
D. Host_name - name of the host
E. Neighbourhood_group - location
F. Neighbourhood - area
G. Latitude - latitude coordinates
H. Longitude - longitude coordinates
I. Room_type - listing space type
J. Price - price in dollar
K. Minimum_nights - amount of nights minimum
L. Number_of_reviews - number of reviews
M. Last_review - latest review
N. Reviews_per_month - number of reviews per month
O. Calculated_host_listings_count - amount of listing per host
P. Availability_365 - number of days when listing is available for booking

3. Data Cleaning and Wrangling

The purpose of the data cleaning and wrangling steps are:

1. To ensure that all features are of the correct data type
2. To ensure missing data are properly imputed
3. To create additional potentially useful features
4. To prepare the dataset for exploratory data analysis (EDA) and statistical analysis

3.1 Data Type Correction

All data types are placed correctly in the dataset. I don't have have to worry about the data types.

3.2 Missing Value Imputation

When I use the “isnull().sum()” method for dataframe, I see that ‘name’, ‘host_name’, ‘last_review’, ‘reviews_per_month’ columns have Nan values.

I checked the relationship between nan value rows and columns.

The same rows in last_review and review_per_month columns are nan values. Also for these rows, number_of_reviews is 0.

Dealing with missing data

The columns which are 'name' and 'host_name' are irrelevant with our further work. We can remove them directly.

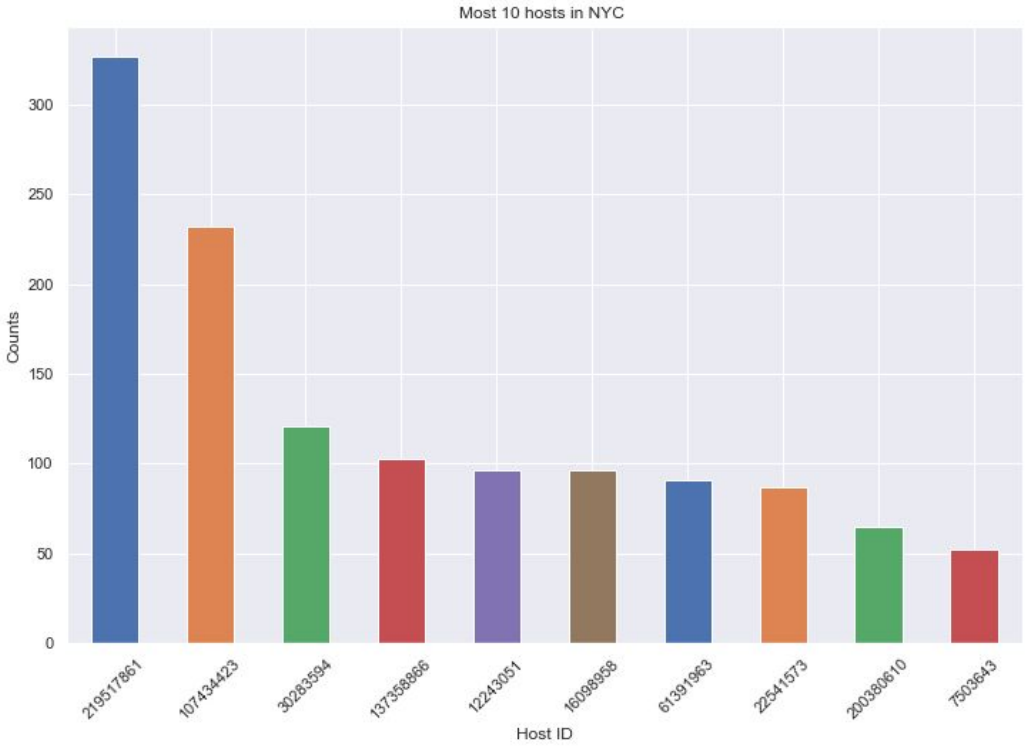
There is no review for the data, then there is no value in ‘last_review’, and ‘review_per_month’ is 0 for these rows of the data. So, we can replace ‘review_per_month’ with 0, and we will remove ‘last_review’ because it is not relevant with our further work.

After we remove 'name' , 'host_name', ‘last_review’ , there are a total 13 columns in the dataset.

4. Data Story

4.1 Host Distribution

We can see that there is a good distribution between top 10 hosts with the most listings. First host has more than 300+ counts and the second host has around 230 counts.

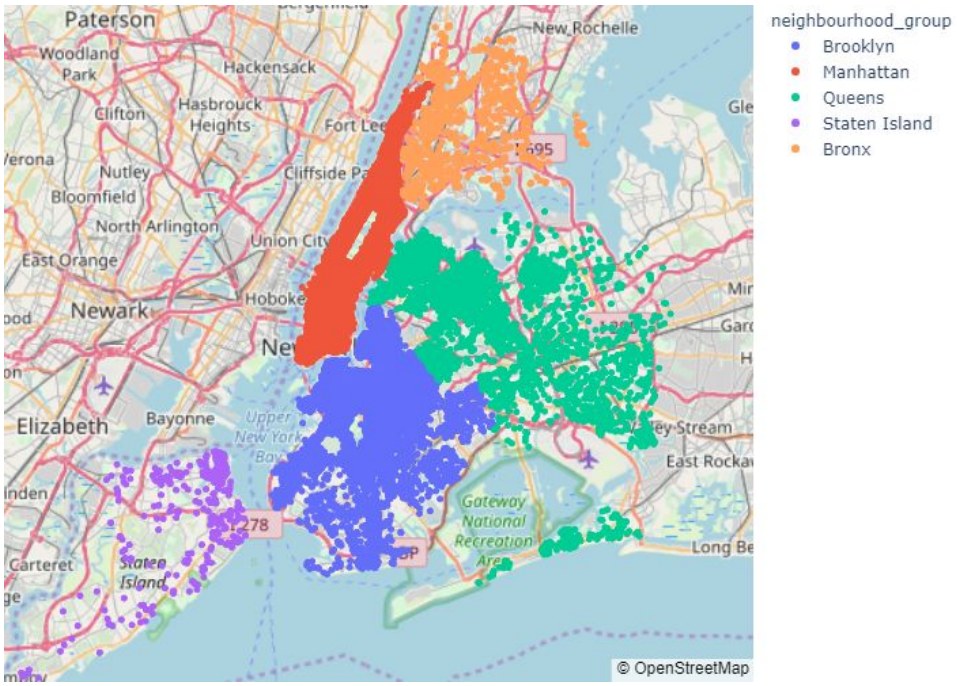


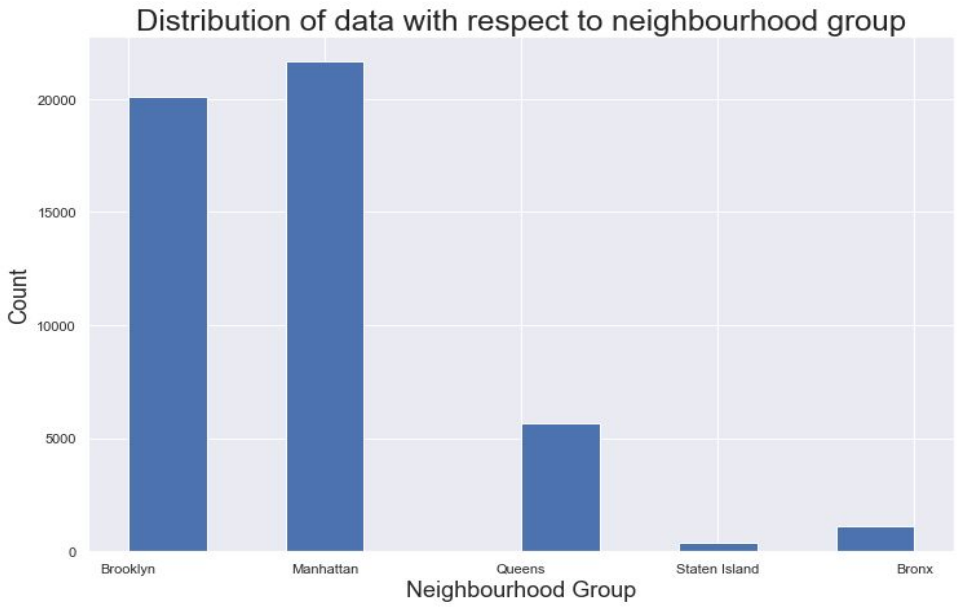
4.2 Price distribution

I notice that most of the prices are distributed between $40-$100 per day. The number of prices are very low above $400. I limited price distribution with $1000 because prices above the $1000 are outliers. The mean price is $133.

4.3 Neighbourhood group

When I check the distribution of neighborhood group numbers, I see that there are about 21000 hosts in Manhattan, 20000 hosts in Brooklyn, 5500 hosts in Queens, 1000 hosts in Bronx and 350 hosts in Staten Island. We can say that most of the hosts in New York City are in Manhattan and Brooklyn. Staten Island and Bronx have very low hosts.

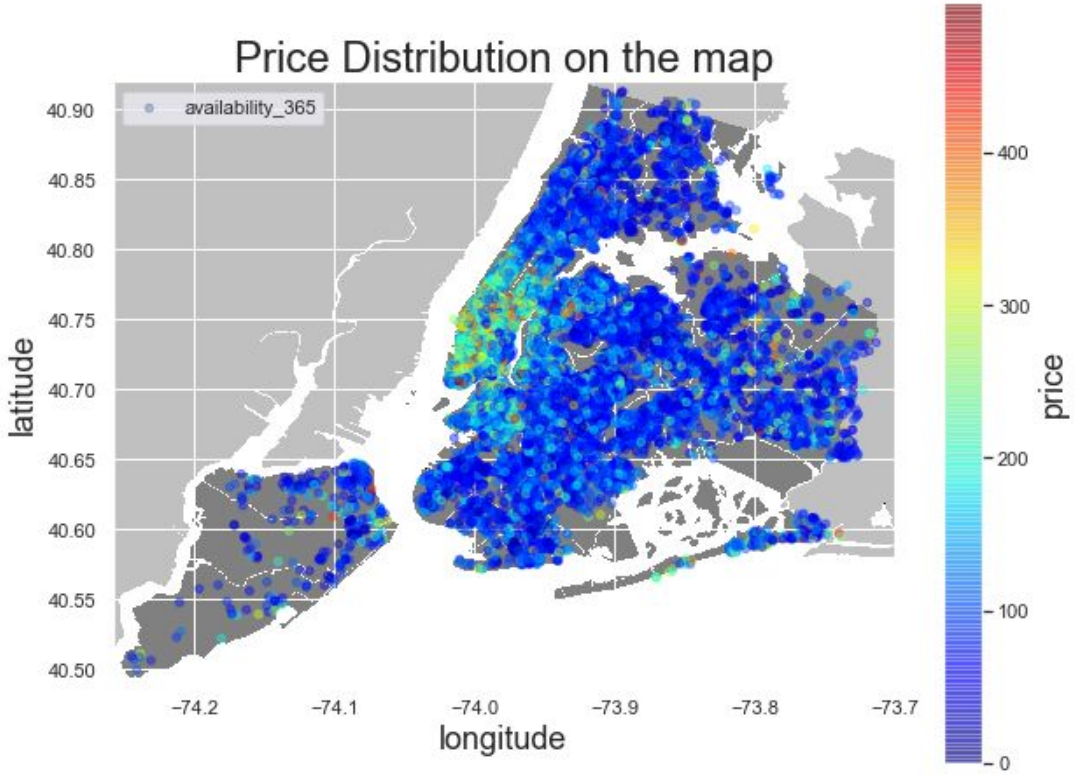




4.3.1 Neighbourhood group and price distribution

Red color dots are the apartment or rooms with higher price. I have considered prices upto 500 $ to get a good representation on the plot.We can see that Manhattan

region has more expensive apartments.



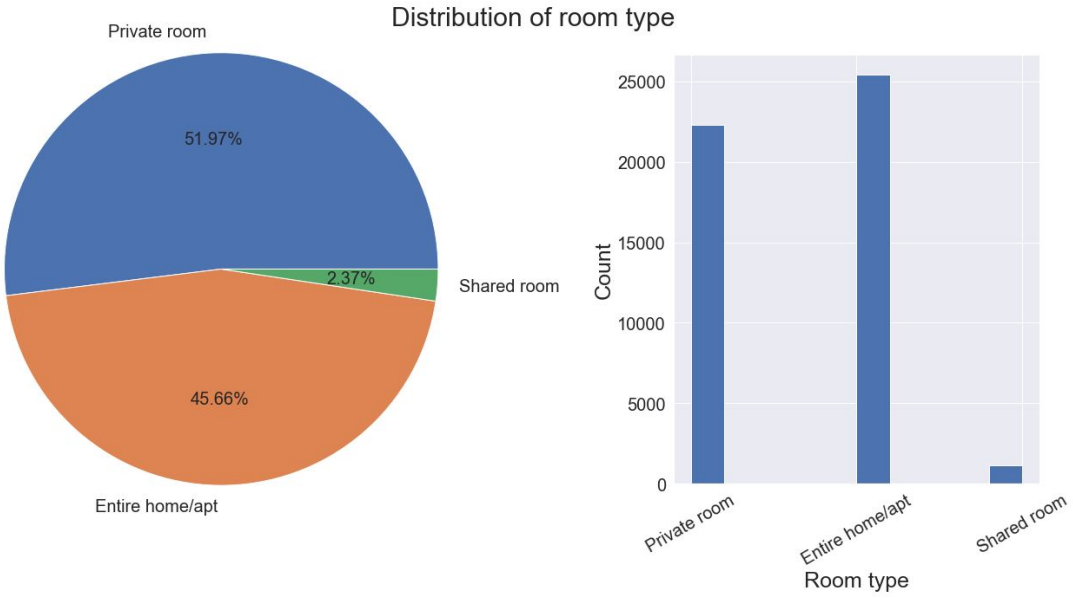
I focused on the mean and range of price distribution between neighbourhood groups. I realize that Manhattan range and mean of price distribution is higher than others. Brooklyn is in second place. Queens and Staten Island appear to have very similar distributions, Bronx is the cheapest of them all

I wonder if these data sets are similar because their hosts counts are very close to each other. I used t test for mean of prices between Manhattan and Brooklyn to see if they are significantly similar. The null hypothesis in these datasets was that both the data sets are significantly similar. The alternative hypothesis was that both the data sets are significantly different. If we consider the significance level to be at 5%, then to accept the null hypothesis, our p-value should be more than the chosen significance level. In the result of the t test, the p-value is 0. Thus we ignore null hypothesis. The t-Test correctly indicates that the means of both the dataset are different and are statistically significantly different from each other.

In the price distribution with respect to neighbourhood groups, Queens and Staten Island's price distribution look similar. To decide if these data sets are significantly similar, I use z-Test for mean of prices between Queens and Staten Island. The null hypothesis in these datasets was that both the data sets are significantly similar. The alternative hypothesis was that both the data sets are significantly different. In the result of the z-Test, the p-value is 0.67. Thus, the p value is greater than my significance level. We accept null hypothesis. The z- Test correctly indicates that the means of both the dataset are statistically significantly similar to each other.

4.4 Room Types

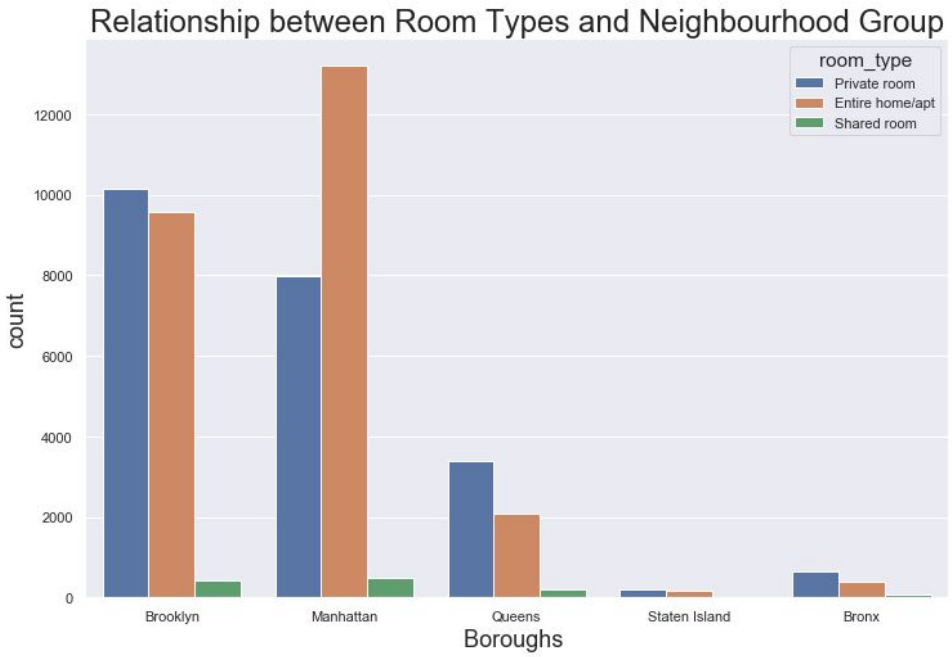
When we check the listing distribution of room types in New York City. Entire home or apartment counts are like 25000, private rooms are about 23000 and shared rooms are like 1160. Counts of private rooms and entire home /apt are very close to each other and shared room count is very low. We can see that Entire home apartment has highest number followed by private room, and least prefered is shared room



4.4.1 Room Type and Neighbourhood Group Distribution

When we check the room type distribution, we can say that Manhattan and Brooklyn have a very high number of listings for each room type. That was somewhat expected as Manhattan and Brooklyn are one of the most traveled destinations, therefore would have the most listing availability. While Manhattan is in the first place at Entire home/apt and shared room, it is in the second place in the private room category.

Queens is in the third place for every room type. Staten Island has a very small number of listings. It’s in the last place for each room type. So we can state that Staten Island is the least traveled destination in New York City.



4.4.2 Room Type and Price Distribution

Mean of price for Entire home/apt is around $150 and it is the highest mean in room types. Also, it has the highest range. The mean of private room price is around $75 and the mean for shared room is $50. I have considered prices upto 600 $ to get a good representation on the violin graft. As we see, their means are not too different, and when we look at the violin graph, their range is very similar. That's why I wonder if they are similar.

I use z-Test for means of private room and shared room. The null hypothesis in these datasets was that private room and shared room are significantly similar. The alternative hypothesis was that both the data sets are significantly different. I consider that the significance level is 5% to accept the null hypothesis, our p-value should be more than the chosen significance level. In my test result, the p-value=0 is well under 0.05 (5%) thus we ignore null hypothesis. The z- Test correctly indicates that both of the datasets are statistically significantly different from each other.

5.Modelling

5.1 Data Preparation for Modelling

Firstly, I found the upper limit of price columns to remove outliers. The upper limit of price distribution is $334. I drop the prices above $334. When I checked the shape of the data, I still had 45918 rows . So, I dropped around 3000 data. It's not a big deal because I still have enough data to decide which model has the best performance in machine learning.

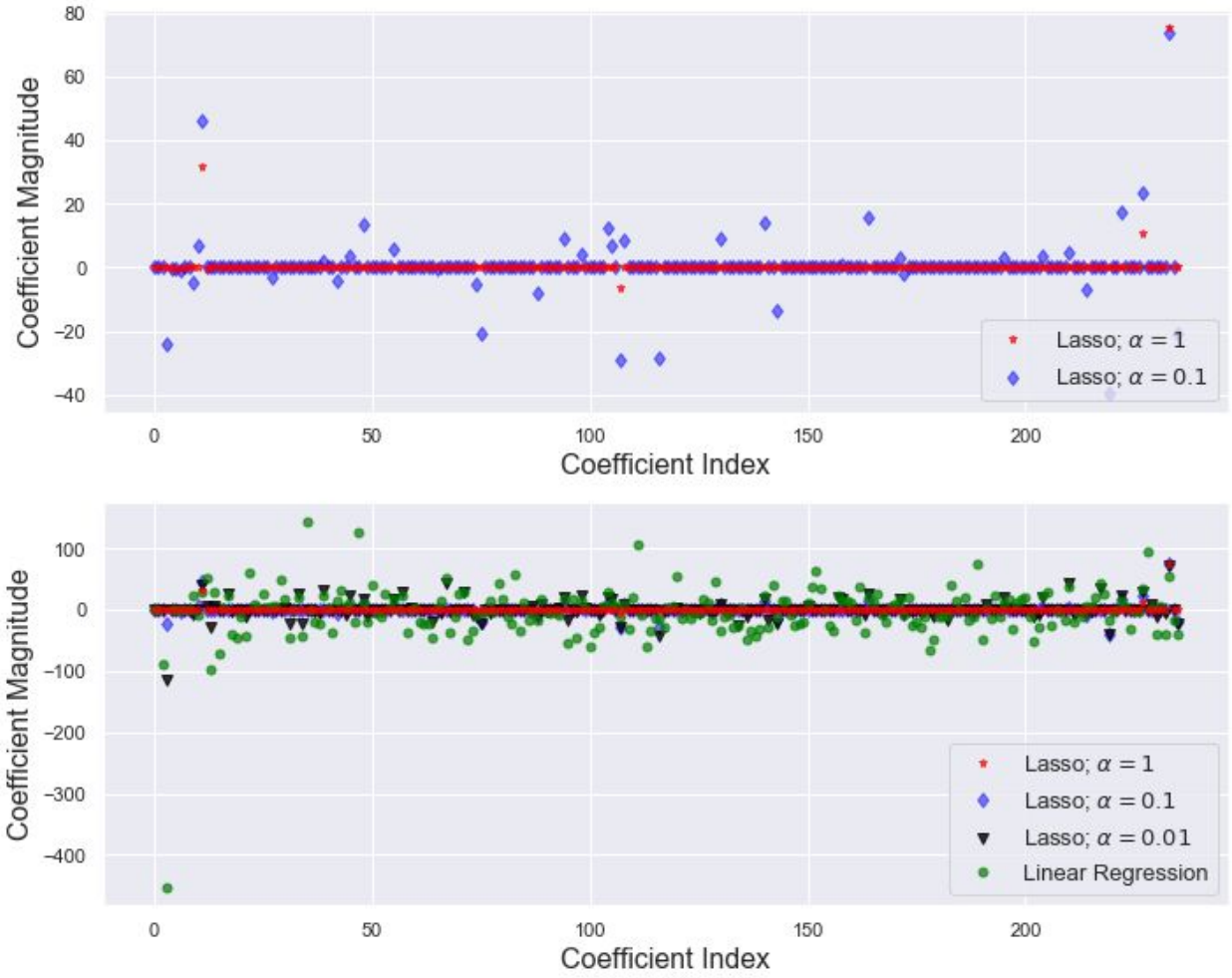
In the data set, types of 3 columns which are 'neighbourhood', neighbourhood_group' and 'room_type' are objects. I converted them to numerical value. I had 237 columns because conversion created columns for each category in the object columns.

I created feature columns that include all columns except price. Price is the target column to make price estimation. And then, I splitted the data into training(%70) and testing samples (%30).

5.2 Regression Models

First of all, I used regression models and I scored R^2 and RMSE(Root Mean Squared Error). **R^2** shows that a statistical measure of how close the data are to the fitted regression line. In general, the higher the R-squared, the better the model fits your data. The **RMSE** is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data–how close the observed data points are to the model's predicted values.

I used Linear Regression with default parameters. R^2 is 0.52 and RMSE is 46. After linear regression I used Lasso Regression with different regularization(alpha) parameters which are 1, 0.1, 0.01. I got the best result alpha=0.01. R^2 is 0.52 and RMSE is 46.6. I visualize them to compare relationships between them.



Let’s understand the plot and the code in a short summary. The default value of the regularization parameter in Lasso regression (given by α) is 1. With this, out of 237 features in the data-set, only 11 features are used (non zero value of the coefficient). Now α = 0.1, non-zero features~32, training and test score increases. Comparison of coefficient magnitude for two different values of alpha are shown in the left panel of figure 2. For alpha =1, we can see most of the coefficients are zero or nearly zero, which is not the case for alpha=0.1. Further reduce α =0.01, non-zero features are increasing. Training and test scores are similar to basic linear regression cases.

In addition, I used Ridge Regression, ElasticNet Regression, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, Linear SVM, Nonlinear SVM.

Model Name	**Best Parameter**	**R^2**	**RSME**
Linear Regression	default	0.526	46.71
Lasso Regression	alpha=0.001	0.524	46.69
Ridge Regression	alpha=1	0.524	46.69
ElasticNet Regression	alpha=0.001	0.523	46.76
Decision Tree Regressor	default	0.16	61.9
Random Forest Regressor	default	0.58	43.88
Gradient Boosting Regressor	default	0.56	44.85
Linear SVM	default	0.0002	135
Nonlinear SVM	default	0.04	70

When we look at the table we can see that the Random Forest Regressor model has the highest R^2 value=0.58 and the lowest RMSE=4385. After I found the best model for the NYC Airbnb dataset, I used the best model for all samples without removing outliers. When I apply the model for all samples R^2 is0.23 and RMSE is 218.

6. Limitations

Data set does not include some essential information like room numbers, bedroom numbers, how many people can live in a house and measures of houses. I believe that these features may impact the result of the price prediction.

7. Conclusion and Next Steps

This Airbnb dataset for the 2019 year is a very rich dataset with a variety of columns. That allowed me to do deep data exploration. Firstly, I explored hosts that provide me host listing distributions on Airbnb. I found that the top host has above 300 listings. Next, I proceed with analyzing price. I saw that most of the prices are distributed between $40-$100 per day and rice counts are very low above $400. My average price is $133. After that, I checked the neighbourhood group and price distribution. We can tell that Brooklyn and Manhattan are the most expensive boroughs of New York City and the most traveled destinations. Lastly, I checked room types and their relationship with price. We saw that the entire home/apt has the highest number followed by the private room, and least prefered is the shared room. Mean of price for Entire home/apt is around $150 and it is the highest mean in room types. Also, it has the highest range. The mean of private room price is around $75 and the mean for shared room is $50. When I applied machine learning with regression models, I got the best result with a random forest regressor. Random forest regression has %58 of variance which is most one. The more variance that is accounted for by the regression model the closer the data points will fall to the fitted regression line.

I couldn’t use different parameters in machine learning which is not possible on my local machine so will have to do that in the cloud.