Data Exploration and Price Prediction Model for Airbnb NYC 2019

Hashneet Kaur

Introduction to Data Science
Final class project report

UCLA Extension – Winter 2020

Contents

1. Project and Dataset Details

Airbnb is a trusted online marketplace for people to list, discover, and book unique accommodations and experiences around the world. Since its launch in 2008, Airbnb has become world's largest community driven hospitality company offering homes & experiences. Whether it is analysis of data for security purposes or to drive more sales, Airbnb has terabytes of data to analyse.

In this project we will study the dataset that lists the details of all the postings on the Airbnb website for New York city in the year 2019.

1.1 Understanding the dataset

The dataset used for this project is taken from the website Kaggle.com. It is available at the below link:

https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data/kernels

The original source of data is the Airbnb website. The dataset provides details about hosts, geographical availability, and necessary metrics to make predictions and draw conclusions. It has 16 different columns.

Detailed description of the dataset using 'Summary'

```
id                                    name          host_id              host_name
Min.   :    2539   Hillside Hotel                :   18   Min.   :     2438   Michael     :  417
1st Qu.: 9471945   Home away from home           :   17   1st Qu.:  7822033   David       :  403
Median :19677284                                 :   16   Median : 30793816   Sonder (NYC):  327
Mean   :19017143   New york Multi-unit building  :   16   Mean   : 67620011   John        :  294
3rd Qu.:29152178   Brooklyn Apartment            :   12   3rd Qu.:107434423   Alex        :  279
Max.   :36487245   Loft Suite @ The Box House Hotel:   11   Max.   :274321313   Blueground  :  232
(Other)                          :48805                    (Other)     :46943
neighbourhood_group           neighbourhood      latitude       longitude              room_type
Bronx        : 1091    Williamsburg      : 3920   Min.   :40.50   Min.   :-74.24   Entire home/apt:25409
Brooklyn     :20104    Bedford-Stuyvesant: 3714   1st Qu.:40.69   1st Qu.:-73.98   Private room   :22326
Manhattan    :21661    Harlem            : 2658   Median :40.72   Median :-73.96   Shared room    : 1160
Queens       : 5666    Bushwick          : 2465   Mean   :40.73   Mean   :-73.95
Staten Island:  373    Upper West Side   : 1971   3rd Qu.:40.76   3rd Qu.:-73.94
Hell's Kitchen    : 1958   Max.   :40.91   Max.   :-73.71
                     (Other)          :32209
    price         minimum_nights    number_of_reviews     last_review    reviews_per_month
 Min.   :    0.0   Min.   :   1.00   Min.   :  0.00             :10052   Min.   : 0.010
1st Qu.:   69.0   1st Qu.:   1.00   1st Qu.:  1.00    2019-06-23: 1413   1st Qu.: 0.190
Median :  106.0   Median :   3.00   Median :  5.00    2019-07-01: 1359   Median : 0.720
Mean   :  152.7   Mean   :   7.03   Mean   : 23.27    2019-06-30: 1341   Mean   : 1.373
3rd Qu.:  175.0   3rd Qu.:   5.00   3rd Qu.: 24.00    2019-06-24:  875   3rd Qu.: 2.020
Max.   :10000.0   Max.   :1250.00   Max.   :629.00    2019-07-07:  718   Max.   :58.500
                                                      (Other)   :33137   NA's   :10052
calculated_host_listings_count availability_365
Min.   :  1.000                Min.   :  0.0
1st Qu.:  1.000                1st Qu.:  0.0
Median :  1.000                Median : 45.0
Mean   :  7.144                Mean   :112.8
3rd Qu.:  2.000                3rd Qu.:227.0
Max.   :327.000                Max.   :365.0
```

Figure 1: Summary of the raw dataset – Airbnb NYC 2019

2. Exploratory Data Analysis

Exploratory data analysis is the initial process of analyzing the data to discover patterns in the variation of different variables, check faulty data and test hypothesis. It is a very critical part of data analysis. The process of EDA is generally done through different types of graphs.

In this project, the dataset has been analyzed based on different variables and their combinations.

2.1. Fixing the missing values

Before we start the process of graphically exploring the data, we must prepare the data. The values that are missing or have NA's must be handled for correct visualization of the data.

From the summary of the dataset we can see that the column 'last_review' has some blank values and and the column number of 'reviews_per_month' has NA's. For our data analysis we can replace NA's in 'reviews_per_month 'with a 0. Since we will not be using columns 'last_review' , 'host_name' and 'ID' for our data analysis, we can remove these columns from our dataframe. Please note that the column 'host_name' is dropped not only because it is insignificant but also for privacy reasons. These are names of actual hosts who have posted listings in Airbnb website.

A look at the final dataset before we start the EDA process:

```
Observations: 48,895
Variables: 13
$ name                            <fct>
$ host_id                         <int>
$ neighbourhood_group             <fct>
$ neighbourhood                   <fct>
$ latitude                        <dbl>
$ longitude                       <dbl>
$ room_type                       <fct>
$ price                           <int>
$ minimum_nights                  <int>
$ number_of_reviews               <int>
$ reviews_per_month               <dbl>
$ calculated_host_listings_count <int>
$ availability_365                <int>
```

Figure 2: Dataset details after data preparation

Description of data types: <fct> - Factor, <int> - Integer, <dbl> - Double

2.2 Exploring the data using exploratory graphs

Now let's plot different variables to explore the data and check our assumptions.

2.2.1 Price density distribution

Let's plot a histogram and analyse the price density distribution for all the listings.

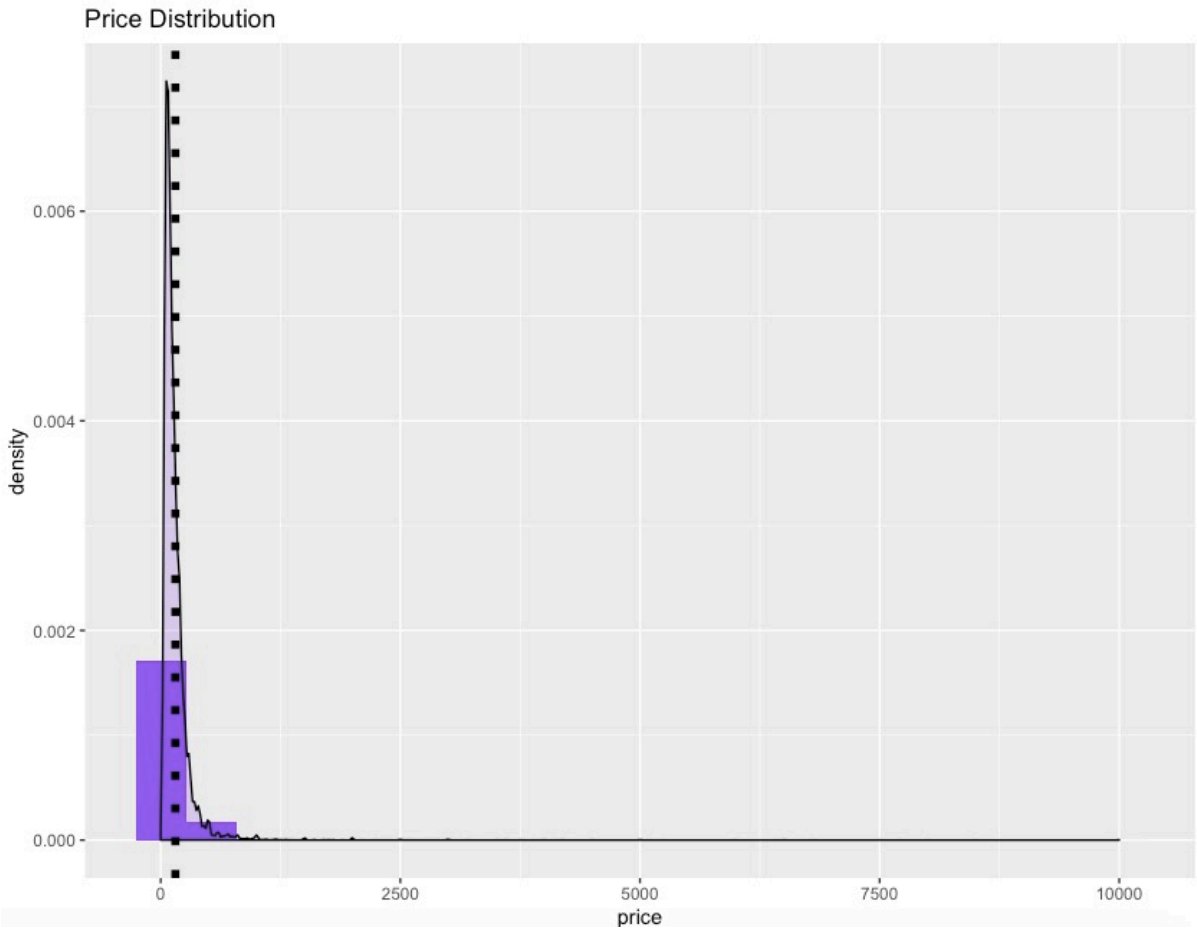


Figure 3: Price density distribution

Since the graph looks very skewed, let's plot this graph again by distributing the price to log 10 for better visualization.

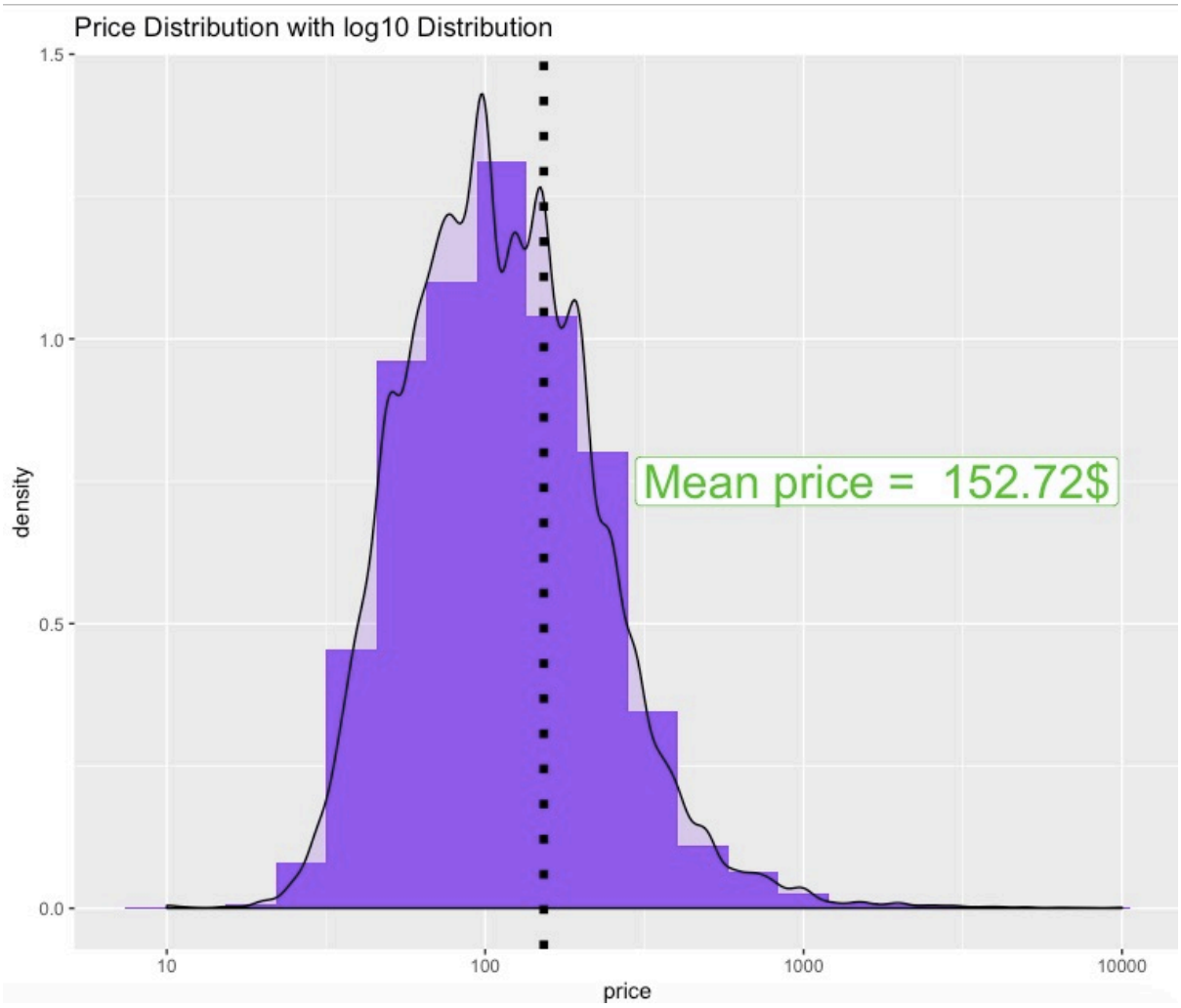


Figure 4: Price density distribution with log 10

The price distribution graph after log 10 distribution is much better and it shows that the mean value of price distribution is $152.72. This price density distribution graph includes all neighbourhood groups. The dataset has 5 different neighbourhood groups:

1. Manhattan
2. Brooklyn
3. Queens
4. The Bronx
5. Staten Island

Let's analyse the mean price for each neighbourhood group by plotting the price density distribution for each one of them.

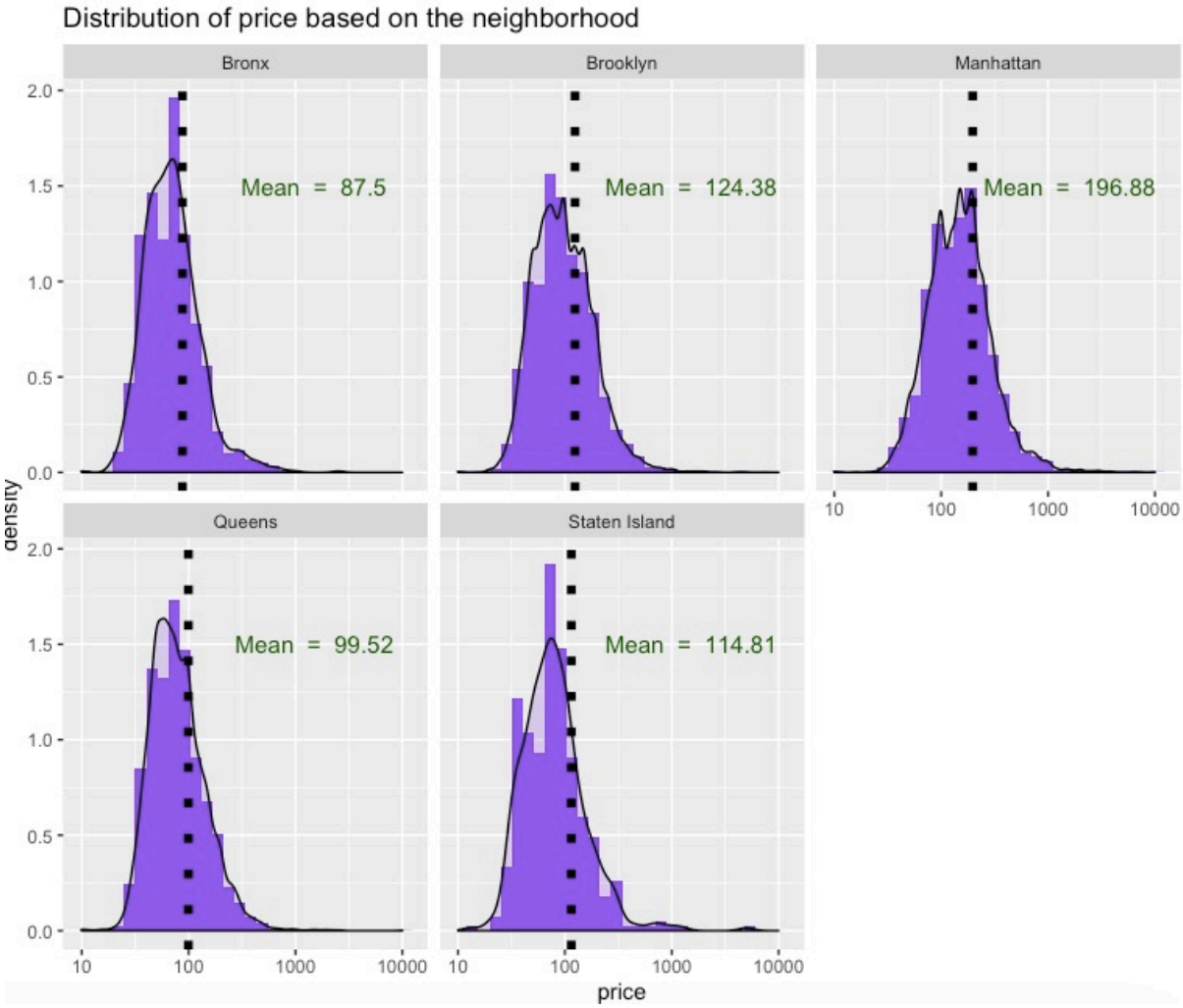


Figure 5: Price density distribution in each neighbourhood

Now let's check out the distribution of prices in each neighbourhood using percentile.

Neighbourhood	***0%***	***25%***	***50%***	***75%***	***100%***
Manhattan	0	95	150	220	10000
Brooklyn	0	45	65	99	2500
Queens	0	45	65	99	2500
The Bronx	0	45	65	99	2500
Staten Island	0	45	65	99	2500

With a statistical table and the histograms we can definitely observe a couple of things about distribution of prices. First, it is visible that Manhattan has the highest range of prices for the listings with $196.88 price as the mean, followed by Brooklyn with $124.28 per night. These two are followed by Staten island, Queens and Bronx in that order. This distribution and density of prices were completely expected; based on the fact that Manhattan is one of the most expensive places in the world to live in, whereas Bronx, on other hand, has the lowest cost of real estate among the five boroughs of NYC.

From the percentile table we can also see that there some listings in the dataset for which the prices are absurdly higher (*the entries with price as 10,000 in Manhattan or with price as 2500 in Brooklyn*). These values are called the outliers.

2.2.2 Room Type and Price Distribution

There are 3 different kinds of room types in the data. These are:

1. Entire home/apt
2. Private room
3. Shared room

Let's now create a box plot for the price distribution based on these different room types.

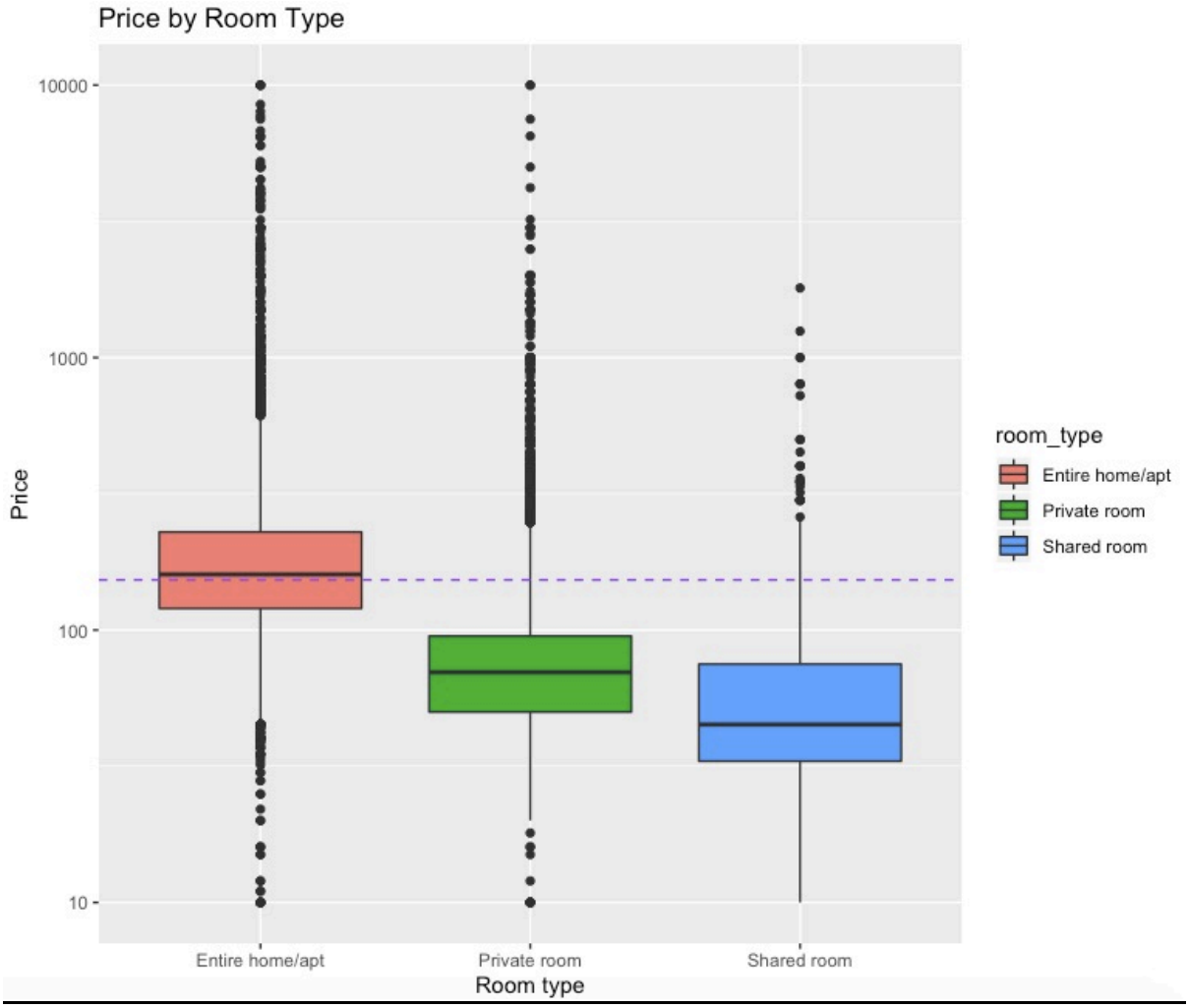


Figure 6: Distribution of price based on room type

As expected ,from these plots it is visible that the highest mean price is for 'Entire home/apt and lowest for shared rooms. Now to study this in detail further let's plot the price distribution based on the room type and neighbourhood group together.

2.2.3 Price distribution based on the room type and neighbourhood group

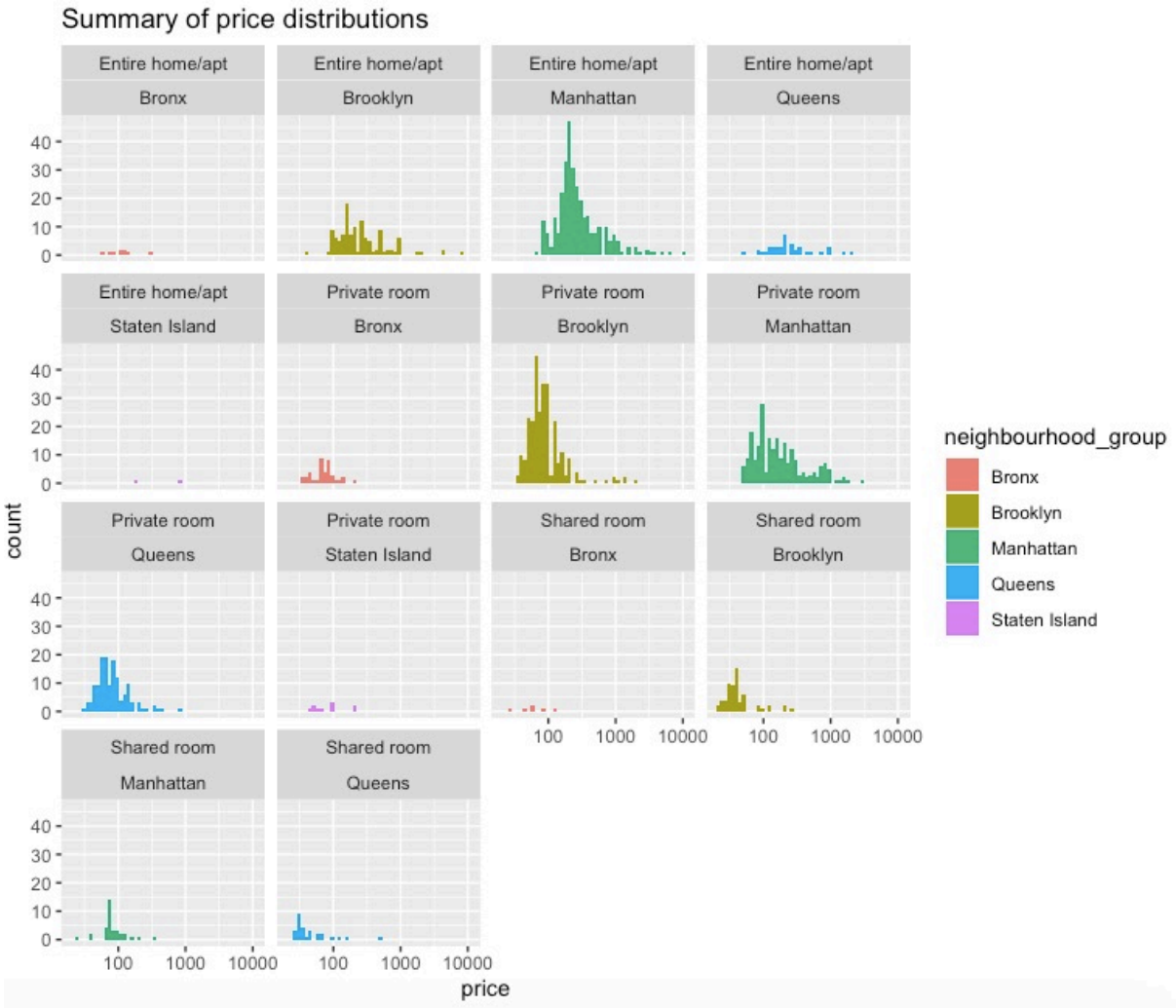


Figure 7: Price, room type and neighbourhood group

These histograms are consistent with the price distribution we saw in figure 6 and figure 5. As we would expect, renting an entire home/apt in Manhattan is the costliest which is followed by renting a similar accommodation in Brooklyn. Another thing that we can observe from these graphs is that the Shared room has minimum number of listings across all neighbourhood groups.

2.2.4 The number of listings in each neighbourhood group

Next let's create a histogram to see the number of listings in each neighbourhood group.

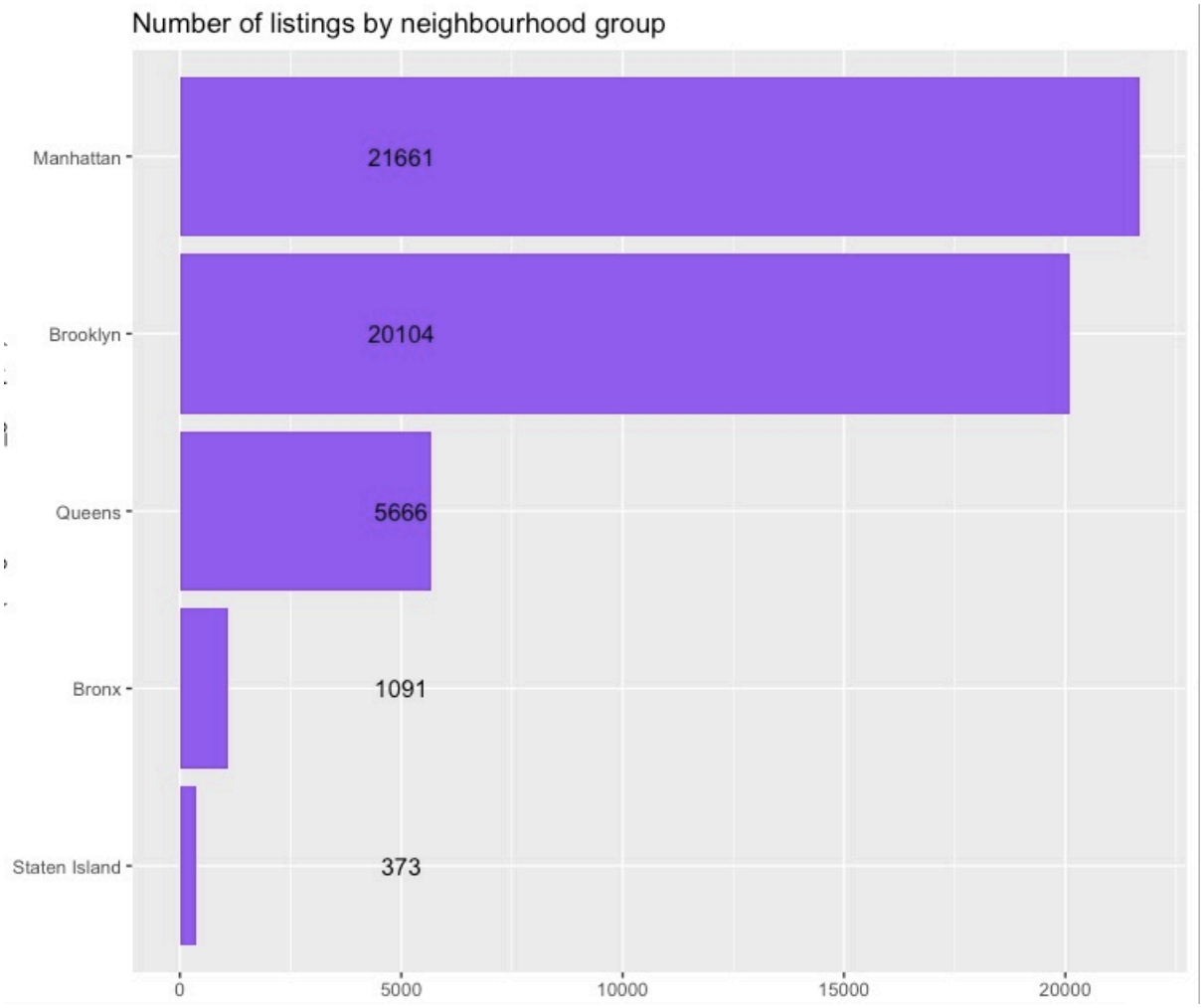


Figure 8: Number of listings in each neighbourhood group

The maximum number of listings are in Manhattan area (21661) and the least are in Staten Island (373). This again seems consistent with what we would expect it to be.

2.2.5 Price and Availability

Let's plot the price distribution against the number of days an accommodation is available for rental through the year.

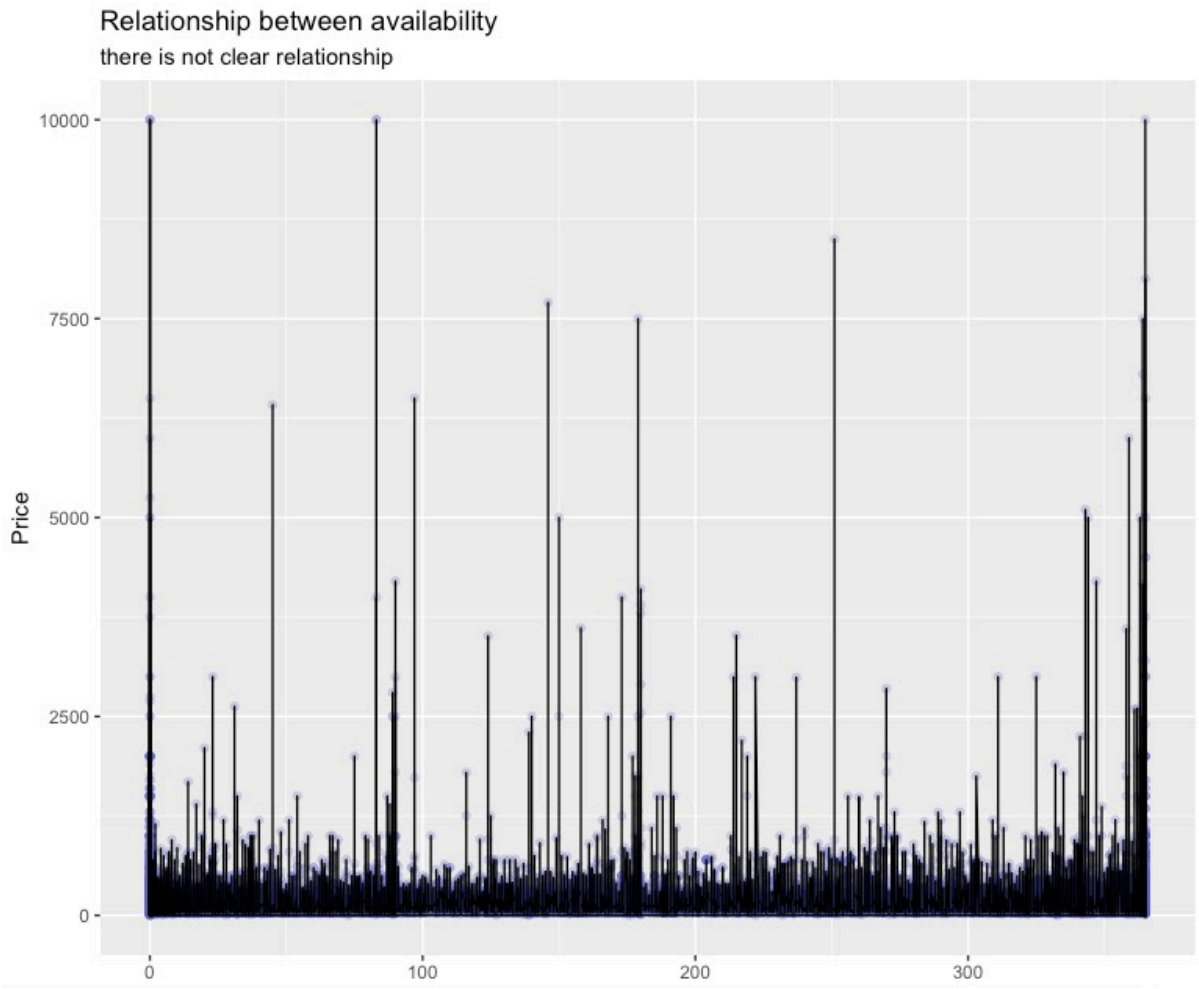


Figure 9: Price and availability

The above graph is very spread out. There is no particular relationship between the price and availability of the listings.

2.2.6 Longitude and Latitude values

Let's create a scatterplot see the different neighbourhood groups based on the longitude and latitude values in the dataset.

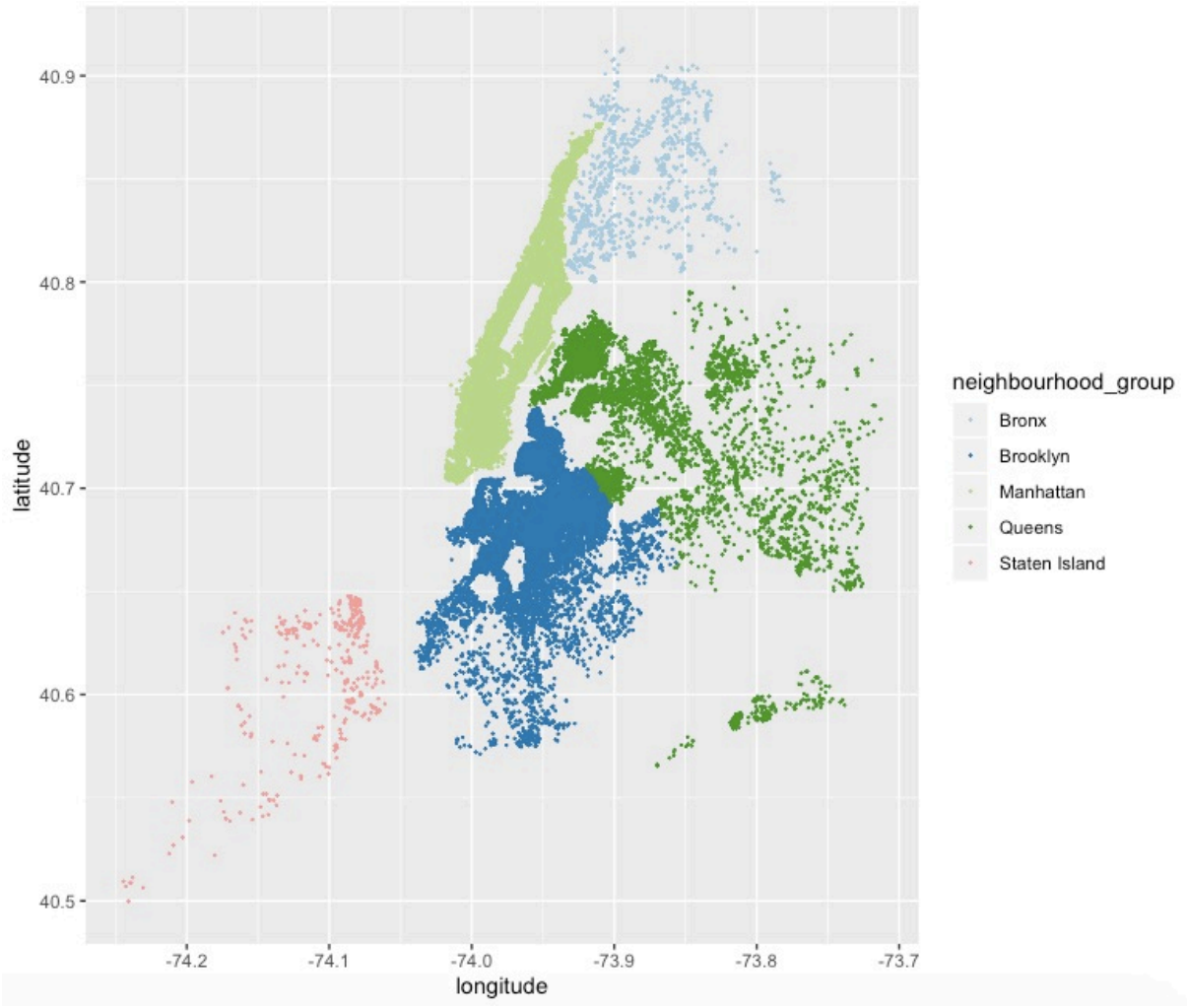


Figure 10: Longitude and Latitude values

2.2.7 Correlation Matrix

A correlation matrix is a table of correlation coefficients for a set of variables. These correlation coefficients are used to determine if a relationship exists between any of those two variables. The coefficient indicates both the strength of the relationship as well as the direction (positive vs. negative correlations)

So next let's plot a correlation matrix for different variables in our dataset.

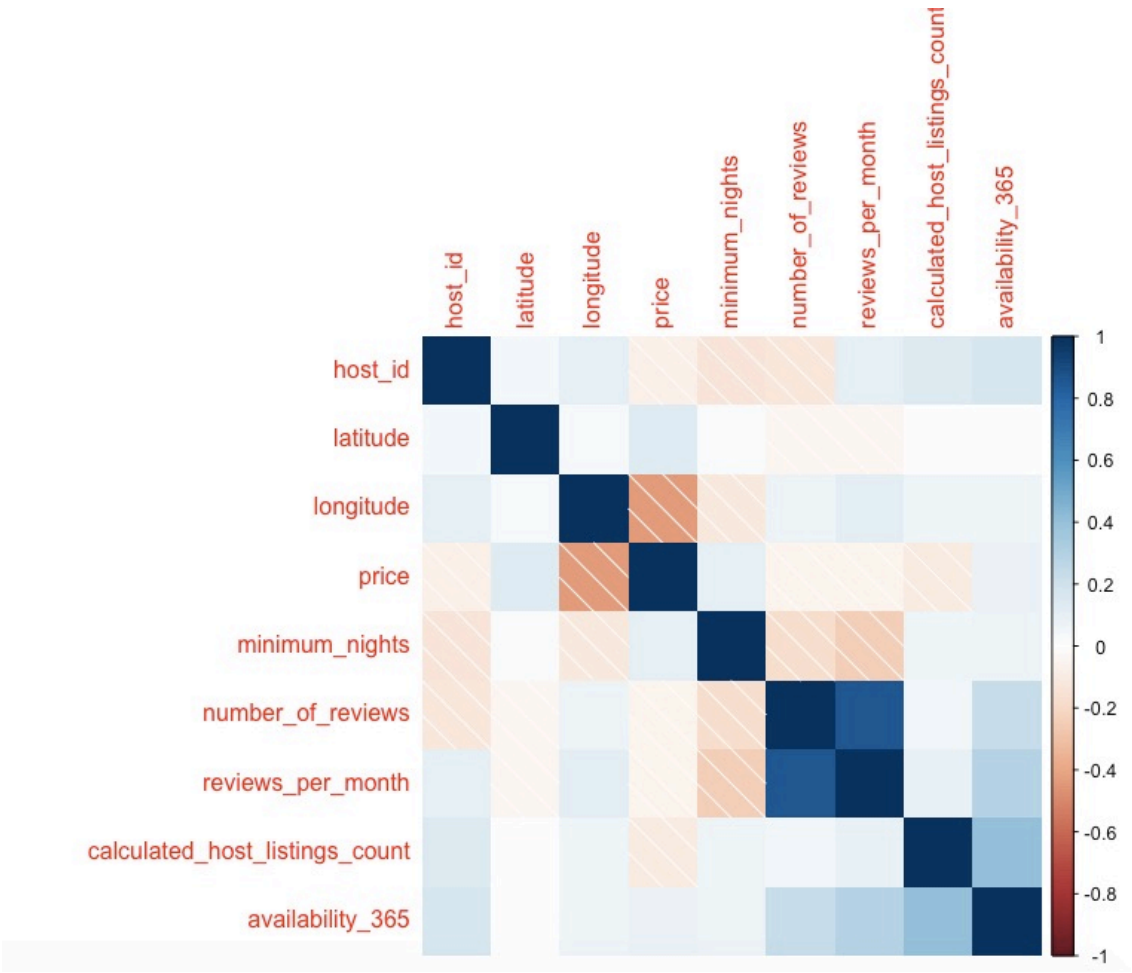


Figure 11: Correlation Matrix

The darker the color the stronger is the co-relation between 2 variables. Also, if the relationship is negative it indicates that if one variable increases, the other will decrease. For example: Price has a negative relationship with number of reviews. The higher the price lesser is the number of reviews for the listing.

3.**Machine learning Algorithm for predicting the price**

Supervised machine learning is the concept of function approximation, where we train an algorithm with test data and in the end of the process we pick the function that best describes the input data, the one that for a given X makes the best estimation of y (X -> y).

3.1 Linear Regression Model 1

There are different types of machine learning algorithms. For this project we will be using linear regression model. This regression model finds a linear relationship between X and y.

Let's create the first linear regression model to predict the price based on the latitude, longitude, room_type, minimum_nights, availability_365 and neighbourhood_group.

```
# Call:
#  lm(formula = price ~ latitude + longitude + room_type + minimum_nights +
#       availability_365 + neighbourhood_group, data = train_abnyc)
#
# Residuals:
#  Min      1Q  Median      3Q     Max
# -192.37  -38.45  -11.20   20.19  417.55
#
# Coefficients:
#  Estimate Std. Error  t value Pr(>|t|)
# (Intercept)                    -2.068e+04  1.193e+03  -17.335  < 2e-16 ***
#  latitude                       -1.104e+02  1.167e+01   -9.461  < 2e-16 ***
#  longitude                      -3.427e+02  1.335e+01  -25.674  < 2e-16 ***
#  room_typePrivate room          -8.668e+01  8.029e-01 -107.966  < 2e-16 ***
#  room_typeShared room           -1.106e+02  2.574e+00  -42.954  < 2e-16 ***
#  minimum_nights                 -1.801e-01  1.832e-02   -9.832  < 2e-16 ***
#  availability_365                8.241e-02  3.024e-03   27.250  < 2e-16 ***
#  neighbourhood_groupBrooklyn    -1.748e+01  3.224e+00   -5.421 5.98e-08 ***
#  neighbourhood_groupManhattan    2.267e+01  2.916e+00    7.775 7.80e-15 ***
#  neighbourhood_groupQueens       7.475e-01  3.106e+00    0.241     0.81
#  neighbourhood_groupStaten Island -1.065e+02  6.160e+00  -17.283  < 2e-16 ***
#  ---
#  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Residual standard error: 65.17 on 28594 degrees of freedom
# Multiple R-squared:  0.4194,  Adjusted R-squared:  0.4192
# F-statistic:  2066 on 10 and 28594 DF,  p-value: < 2.2e-16
```

For a linear regression model to be a good fit the median residual error has to be close to 0. And the R-squared value has to be close to 1. This linear regression model is not a good model as the median residual error is -11.20 which is very far away from 0. The value of R-squared is also not good as it is less than 0.5

3.2 Linear Regression 2

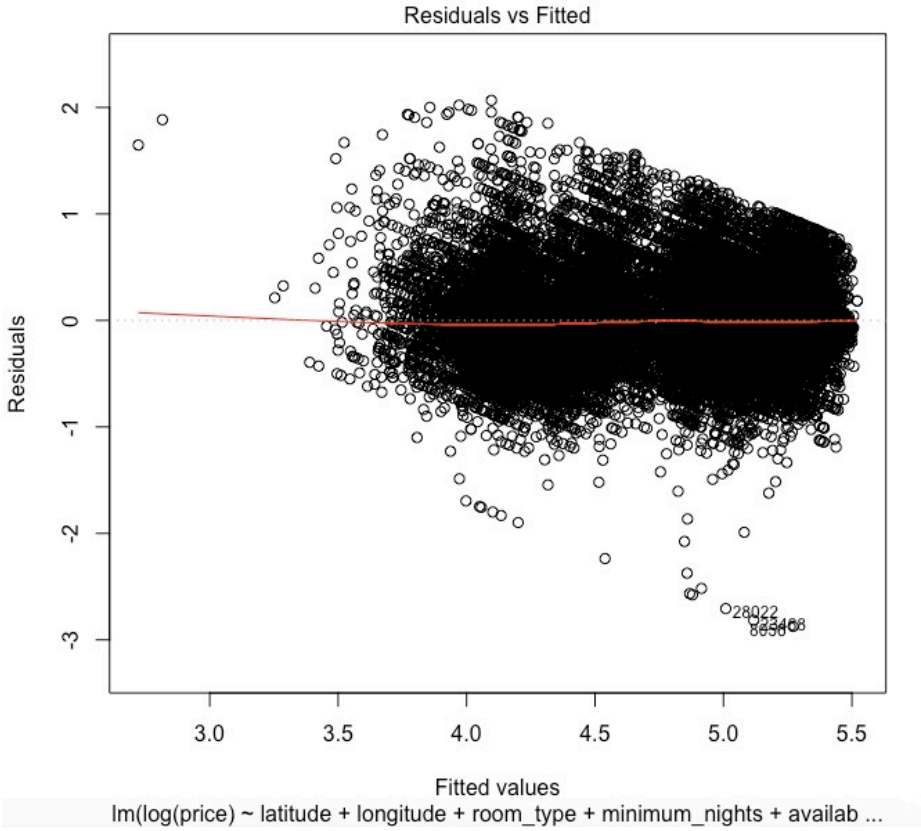
Since the previous linear regression model was not a good model, let's create another linear regression model by log 10 conversion of the price. We will base the model on the same set of variables.

```
# Call:
#  lm(formula = log(price) ~ latitude + longitude + room_type +
#       minimum_nights + availability_365 + neighbourhood_group,
#      data = train_abnyc)

# Residuals:
#  Min       1Q   Median       3Q      Max
# -2.87348 -0.28066 -0.02938  0.24789  2.06662

# Coefficients:
#  Estimate Std. Error  t value Pr(>|t|)
# (Intercept)                    -1.787e+02  7.768e+00  -23.009  < 2e-16 ***
#  latitude                       -5.078e-01  7.596e-02   -6.685 2.35e-11 ***
#  longitude                      -2.763e+00  8.691e-02  -31.790  < 2e-16 ***
#  room_typePrivate room          -7.230e-01  5.227e-03 -138.303  < 2e-16 ***
#  room_typeShared room           -1.101e+00  1.676e-02  -65.694  < 2e-16 ***
#  minimum_nights                 -1.550e-03  1.193e-04  -12.998  < 2e-16 ***
#  availability_365                5.429e-04  1.969e-05   27.571  < 2e-16 ***
#  neighbourhood_groupBrooklyn    -2.568e-02  2.099e-02   -1.224    0.221
#  neighbourhood_groupManhattan    2.563e-01  1.898e-02   13.501  < 2e-16 ***
#  neighbourhood_groupQueens       8.828e-02  2.022e-02    4.366 1.27e-05 ***
#  neighbourhood_groupStaten Island -7.267e-01  4.011e-02  -18.118  < 2e-16 ***
#  ---
#  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# Residual standard error: 0.4243 on 28594 degrees of freedom
# Multiple R-squared:  0.5387,  Adjusted R-squared:  0.5385
# F-statistic:  3339 on 10 and 28594 DF,  p-value: < 2.2e-16
```

This model is much better than the previous model as the median residual error is -0.02938 which is really close to 0. Also, the R-squared value is 0.5387 which is better than the previous model. Next, let’s plot and see the different regression graphs

Residuals vs Fitted graph

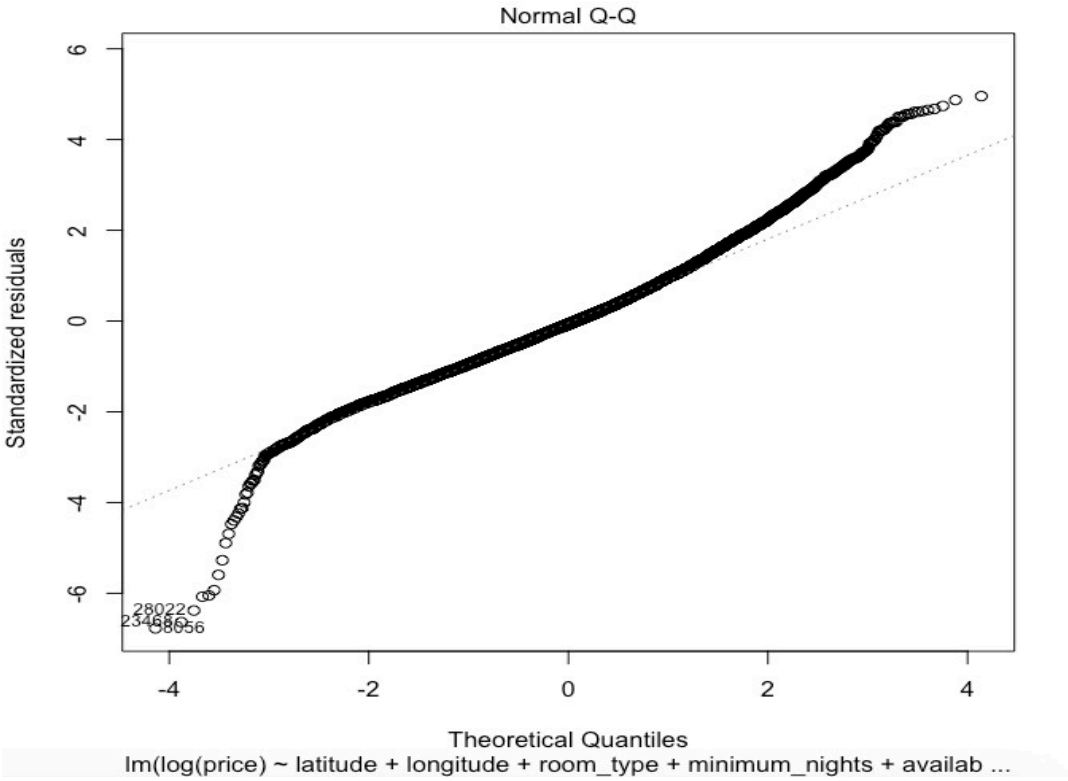
The “Residuals vs. Fitted” plot shows if residuals have non-linear patterns. If the residual values are spread uniformly around a horizontal line without distinct patterns, that is a good indication of a strong relationships.



This residuals vs fitted graph shows a good linear relationship.

Normal Q-Q Plot

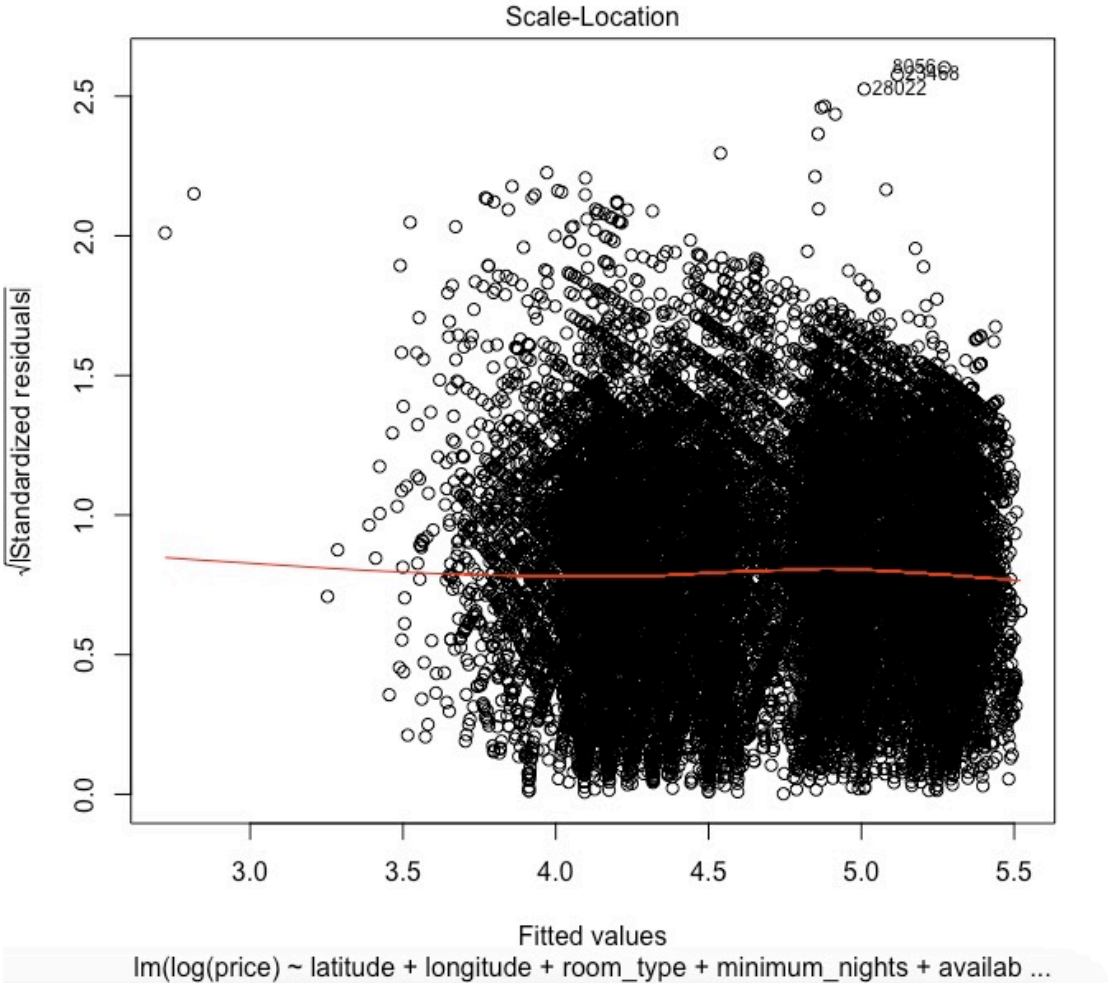
Normal Q-Q plot shows if residuals are normally distributed. The more the Q-Q distribution is linearly distributed along a line, the better the model is.



The Normal Q-Q graph has most of the values spread along the linear regression line with some outliers.

Scale – Location Graph

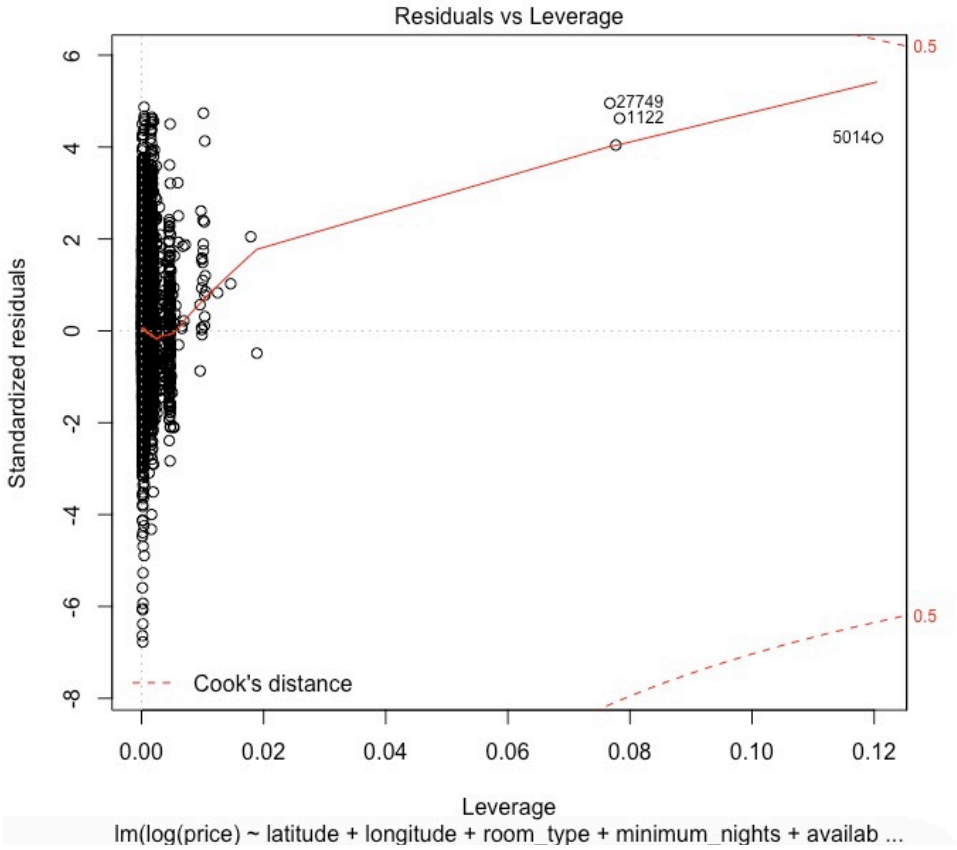
The Scale-Location plot displays the way the residuals are spread along the ranges of predictors. The more gradually the residuals are spread along the predictors the better it is. If the plot shows equal distribution along a horizontal line the better the model is.



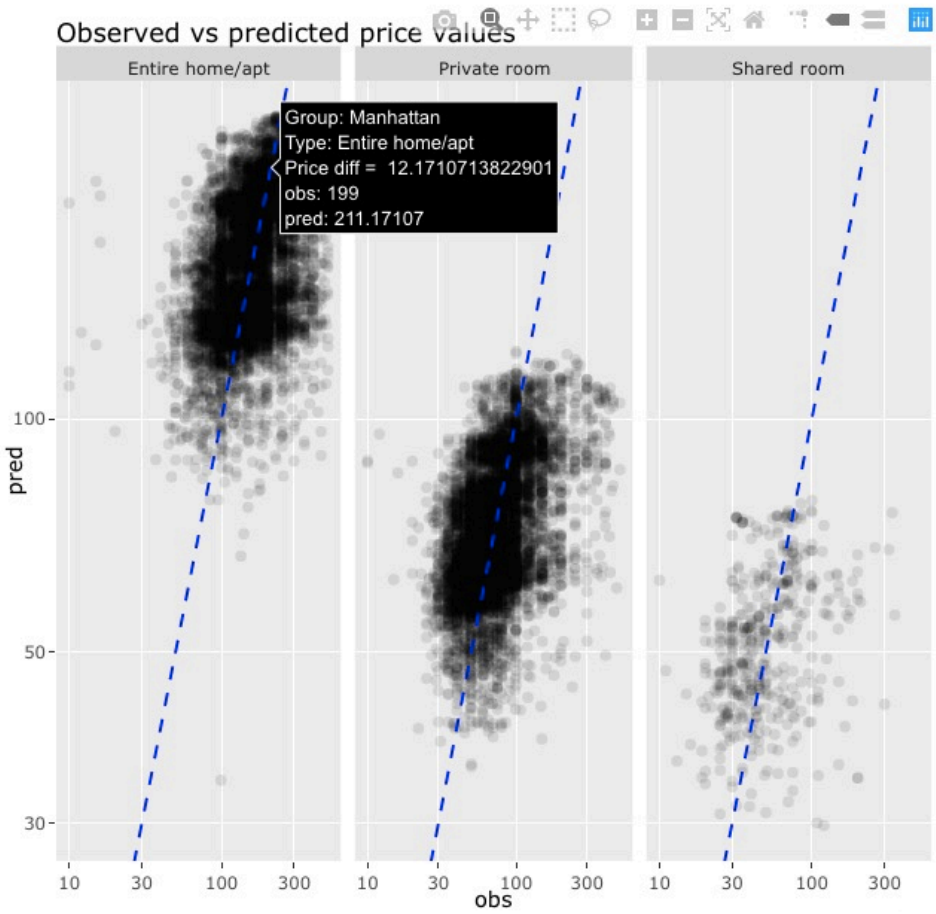
The residuals in this graph are almost gradually distributed along a straight line with some outliers.

Residuals vs Leverage

The Residuals vs. Leverage plots helps to identify influential data points in the model. Outlier values can be influential in some cases. The points that we have to watch out for are values in the upper right or lower right corners, which are outside the red dashed Cook's distance line. These are points that would be influential in the model and removing them would be likely noticeably alter the regression results.

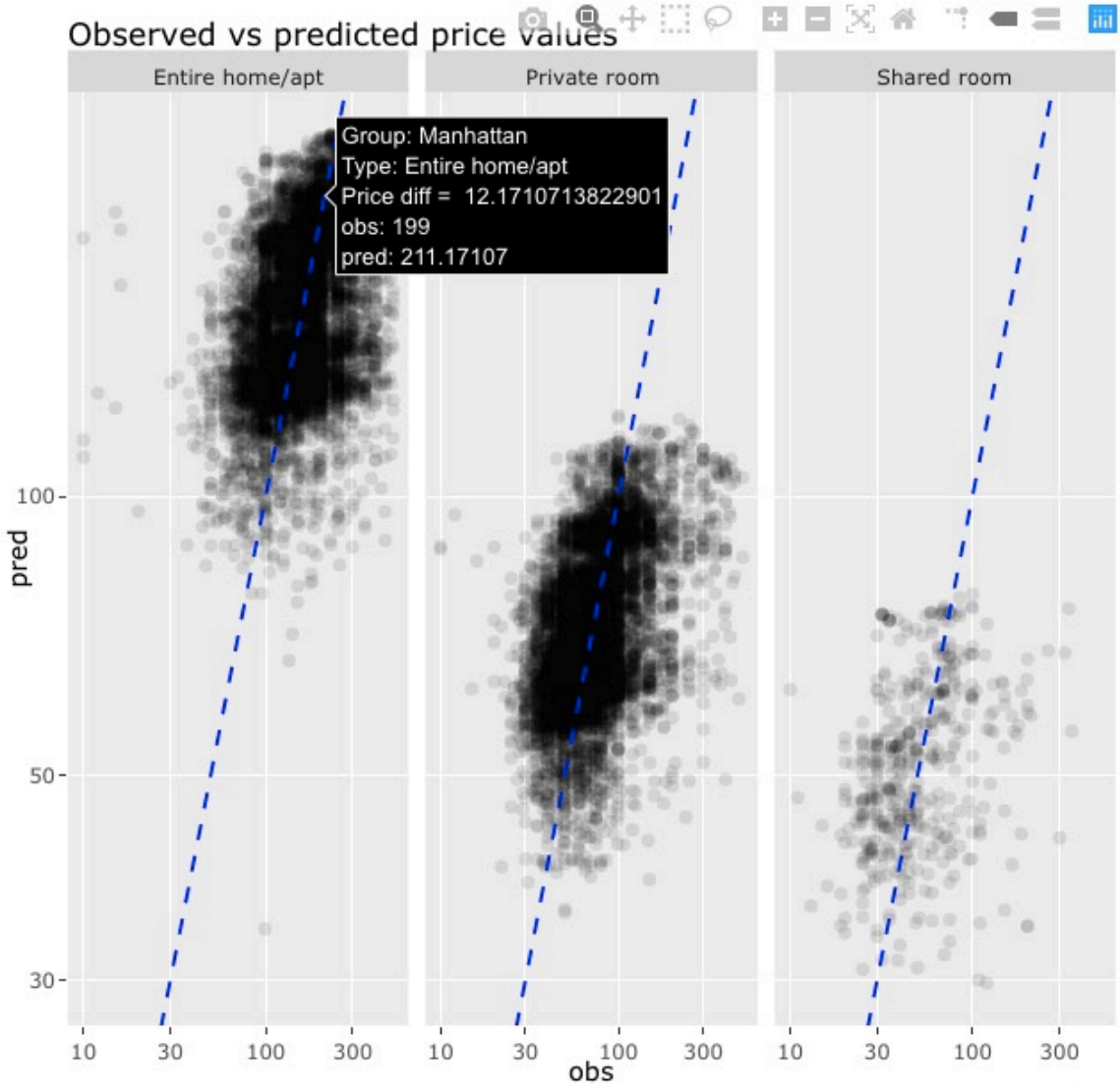


The final step is to verify this model on the test data to check the correctness of the results generated by the model. The aim is of this step is to feed the model with the data for which the answers are known and then compare the result from model with these known values.



4. Conclusion

The final model created can be used to predict the price for Airbnb NYC accomodations by inputting the values of latitude, longitude, room_type, minimum_nights, availability_365 and neighbourhood_group. The predicted value would have an approximate residual error of 40%.


Observed vs predicted price values
Entire home/apt
Private room
Shared room
Group: Manhattan
Type: Entire home/apt
Price diff = 12.1710713822901
obs: 199
pred: 211.17107
pred
100
50
30
10
30
100
300
obs