Business Analytics Tools II

Final Project

New York City Airbnb Data2019

Submitted by:

Divya Bhattiprolu

Dayana Sunil

Niharika Nagpal

Submitted to:

Prof. Jin Man Lee

Introduction

Business Motivation

In recent years, sharing economy has flourished around the world, and the great popularity of smartphones and mobile internet accelerated the emergence and expansion of various forms, such as online car rental, shared bicycle rental, short-term house rental, and the gig economy. Airbnb is a great example of online marketplaces and the sharing economy. With the rise of Airbnb, the short-term house rentals have grown rapidly. The growth of short-term rentals helps tourists experience local life, understand local culture and create employment opportunities for local residents.

Business Question

For this project, we are acting as business consultants who help the new hosts become successful. In broader sense, our business question is:

What makes a host successful?

- *What factors influences a host to become a super host?*
- *What factors influence better review score?*

Note: A host becomes a superhost when: Overall ratings are 4.8+ , there is a 90% response rate within 24 hours , the host has at least 10 stays a year, he/she honors confirmed reservations with no cancellations ,they maintain at least a 50% review rate.

How will we answer our question?

We are describing the data using descriptive statistics. This includes summary statistics based on cluster grouping and central trend statistics for the data. We will use different prescriptive analytic methods such a linear probability, logistic regression, random forest models, and neural networks and pick the best performing model to predict the trend we observe associated with the variables that are chosen.

Data and Empirical Methodology

Data

We are using the New York City Airbnb data 2019 and we have taken it from insideairbnb.com.

The entire dataset had about 70 variables. After looking at the summary of the data, we decided to take the following 32 variable which are required for answering our business questions

```
 $ host_id                       : num [1:44666] 2845 4869 7356 7378 8967 ...
 $ host_since                    : chr [1:44666] "9/9/2008" "12/7/2008" "2/3/2009" "2/3/2009" .
 $ host_response_time            : chr [1:44666] "within a day" "within an hour" "N/A" "within
 $ host_response_rate            : chr [1:44666] "70%" "98%" "N/A" "100%" ...
 $ host_acceptance_rate          : chr [1:44666] "25%" "96%" "100%" "N/A" ...
 $ host_is_superhost             : logi [1:44666] FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ host_listings_count           : num [1:44666] 6 1 1 1 1 1 1 3 2 1 ...
 $ neighbourhood_cleansed        : chr [1:44666] "Midtown" "Clinton Hill" "Bedford-Stuyvesant"
 $ neighbourhood_group_cleansed  : chr [1:44666] "Manhattan" "Brooklyn" "Brooklyn" "Brooklyn" .
 $ latitude                      : num [1:44666] 40.8 40.7 40.7 40.7 40.8 ...
 $ longitude                     : num [1:44666] -74 -74 -74 -74 -74 ...
 $ property_type                 : chr [1:44666] "Entire apartment" "Entire guest suite" "Priva
 $ room_type                     : chr [1:44666] "Entire home/apt" "Entire home/apt" "Private r
 $ accommodates                  : num [1:44666] 2 3 2 4 2 1 2 2 1 3 ...
 $ bedrooms                      : num [1:44666] NA 1 1 2 1 1 1 1 1 NA ...
 $ beds                          : num [1:44666] 1 3 1 2 1 1 1 0 1 1 ...
 $ amenities                     : chr [1:44666] "[\"Hot water\", \"Stove\", \"Extra pillows ar
_ "[\"Hot water\", \"Stove\", \"Free parking on premises\", \"Extra pillows and blankets\", \"S
\", \"Kitchen\"]" "[\"Wifi\", \"Dryer\", \"Air conditioning\", \"Kitchen\", \"Smoke alarm\", \"
 $ price                         : chr [1:44666] "$175.00" "$76.00" "$60.00" "$175.00" ...
 $ minimum_nights                : num [1:44666] 3 1 29 7 2 2 3 4 2 30 ...
 $ maximum_nights                : num [1:44666] 1125 730 730 1125 14 ...
 $ availability_365              : num [1:44666] 365 2 2 359 350 0 125 0 365 0 ...
 $ number_of_reviews             : num [1:44666] 48 354 50 1 473 118 66 181 123 181 ...
 $ review_scores_rating          : num [1:44666] 94 89 90 97 84 98 97 94 93 91 ...
 $ review_scores_accuracy        : num [1:44666] 9 8 8 10 9 10 10 10 9 9 ...
 $ review_scores_cleanliness     : num [1:44666] 9 9 8 10 7 10 9 10 10 10 ...
 $ review_scores_checkin         : num [1:44666] 10 9 10 10 9 10 10 10 10 10 ...
 $ review_scores_communication   : num [1:44666] 10 9 10 10 9 10 10 10 9 10 ...
 $ review_scores_location        : num [1:44666] 10 9 9 8 10 10 10 10 9 9 ...
 $ review_scores_value           : num [1:44666] 9 9 9 10 9 10 10 10 9 9 ...
 $ instant_bookable              : logi [1:44666] FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ calculated_host_listings_count: num [1:44666] 2 1 1 1 1 1 1 3 1 1 ...
```

Data Summary

```
> summary(airbnb1)
 description           host_id            host_since        host_response_time
 Length:44666       Min.   :     2438   Length:44666       Length:44666
 Class :character   1st Qu.:  9354841   Class :character   Class :character
 Mode  :character   Median : 37854928   Mode  :character   Mode  :character
                    Mean   : 85318403
                    3rd Qu.:137358866
                    Max.   :370512147

 host_response_rate host_acceptance_rate host_is_superhost host_listings_count
 Length:44666       Length:44666         Mode :logical     Min.   :   0.00
 Class :character   Class :character     FALSE:36633       1st Qu.:   1.00
 Mode  :character   Mode  :character     TRUE :8016        Median :   1.00
                                         NA's :17          Mean   :  15.82
                                                           3rd Qu.:   2.00
                                                           Max.   :2114.00
                                                           NA's   :17
 neighbourhood_cleansed neighbourhood_group_cleansed    latitude       longitude
 Length:44666           Length:44666                 Min.   :40.51   Min.   :-74.24
 Class :character       Class :character             1st Qu.:40.69   1st Qu.:-73.98
 Mode  :character       Mode  :character             Median :40.72   Median :-73.96
                                                     Mean   :40.73   Mean   :-73.95
                                                     3rd Qu.:40.76   3rd Qu.:-73.93
                                                     Max.   :40.91   Max.   :-73.71


 property_type       room_type          accommodates      bedrooms          beds
 Length:44666       Length:44666       Min.   : 0.000   Min.   : 1.000   Min.   : 0.000
 Class :character   Class :character   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000
 Mode  :character   Mode  :character   Median : 2.000   Median : 1.000   Median : 1.000
                                       Mean   : 2.758   Mean   : 1.292   Mean   : 1.507
                                       3rd Qu.: 4.000   3rd Qu.: 1.000   3rd Qu.: 2.000
                                       Max.   :16.000   Max.   :21.000   Max.   :40.000
                                                        NA's   :4259     NA's   :513
  amenities            price          minimum_nights     maximum_nights
 Length:44666       Length:44666       Min.   :   1.000   Min.   :1.000e+00
 Class :character   Class :character   1st Qu.:   2.000   1st Qu.:2.900e+01
 Mode  :character   Mode  :character   Median :   3.000   Median :3.650e+02
                                       Mean   :   9.301   Mean   :4.957e+04
                                       3rd Qu.:   7.000   3rd Qu.:1.125e+03
                                       Max.   :1250.000   Max.   :2.147e+09

 availability_365 number_of_reviews review_scores_rating review_scores_accuracy
 Min.   :  0.0    Min.   :  0.00    Min.   : 20.00       Min.   : 2.0
 1st Qu.:  0.0    1st Qu.:  1.00    1st Qu.: 92.00       1st Qu.: 9.0
 Median :  1.0    Median :  4.00    Median : 97.00       Median :10.0
 Mean   :105.4    Mean   : 22.46    Mean   : 93.85       Mean   : 9.6
 3rd Qu.:189.0    3rd Qu.: 20.00    3rd Qu.:100.00       3rd Qu.:10.0
 Max.   :365.0    Max.   :746.00    Max.   :100.00       Max.   :10.0
                                    NA's   :11564        NA's   :11596
```

```
review_scores_cleanliness review_scores_checkin review_scores_communication
Min.   : 2.000            Min.   : 2.000        Min.   : 2.000
1st Qu.: 9.000            1st Qu.:10.000        1st Qu.:10.000
Median :10.000            Median :10.000        Median :10.000
Mean   : 9.269            Mean   : 9.724        Mean   : 9.727
3rd Qu.:10.000            3rd Qu.:10.000        3rd Qu.:10.000
Max.   :10.000            Max.   :10.000        Max.   :10.000
NA's   :11583             NA's   :11611         NA's   :11591
review_scores_location review_scores_value instant_bookable calculated_host_listings_count
Min.   : 2.000         Min.   : 2.000      Mode :logical    Min.   :  1.000
1st Qu.: 9.000         1st Qu.: 9.000      FALSE:29619      1st Qu.:  1.000
Median :10.000         Median :10.000      TRUE :15047      Median :  1.000
Mean   : 9.594         Mean   : 9.386                       Mean   :  6.668
3rd Qu.:10.000         3rd Qu.:10.000                       3rd Qu.:  2.000
Max.   :10.000         Max.   :10.000                       Max.   :239.000
NA's   :11616          NA's   :11615
```

We also cleaned the data since it had more than 80,000 N/A values.

Empirical Methodology

Linear Regression- Review scores rating

Formally, the model for multiple linear regression, given *n* observations, is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots \beta_p x_{ip} + \varepsilon_i \text{ for } i = 1,2, \dots n.$$

```
(review_scores_rating ~ room_type +
   number_of_reviews + price + review_scores_accuracy +
   review_scores_cleanliness + review_scores_checkin + review_scores_communication +
   review_scores_location + review_scores_value, availability_365
```

The model explains the relationship between multiple explanatory variables and response variable by fitting a linear equation to observed data. Our dependent variable review score rating is best described by the following independent variables:

Number of reviews, Price, Review scores accuracy, Review score cleanliness, Review scores checkin, review scores communication, review scores location, review scores value and availability 365.

Our model is trying to figure out what factors influence the review score ratings.

Random forest- Superhost

```
rf = randomForest(superhost~., data = a1, ntree = 100)
```

Our dependent variable superhost is best described by the following independent variables:

accommodates, beds, number of reviews, reviews score rating, reviews for cleanliness, check-in, communication, location, value, calculated host listings count.

How will this methodology help us answer the questions?

In order to answer the question, we performed several descriptive and predictive analysis. We cleaned our data because it had many N/A values and created subsets for our dataset. A new variable called superhost was created, which had a numeric value 1/0 as opposed to the original T/F value. Using all the models we learned in class, we concluded that the best model is the Random Forest model since it had the highest AUC.

Results

Descriptive Analytics

FREQUENCIES

	Frequency <dbl>	Percent <dbl>
Staten Island	228	0.7759325
Bronx	790	2.6885380
Queens	3852	13.1091751
Manhattan	11895	40.4812143
Brooklyn	12619	42.9451402

	Frequency <dbl>	Percent <dbl>
Hotel room	245	0.8337871
Shared room	620	2.1099918
Private room	13946	47.4612034
Entire home/apt	14573	49.5950177

MEAN MEDIAN MODE

	Mean <dbl>
Price	140.2405779
Review Scores Rating	0.5901665

	Median <dbl>
Price	99.00000
Review Scores Rating	0.60206

	Mode <fctr>
Price	150
Review Scores Rating	0

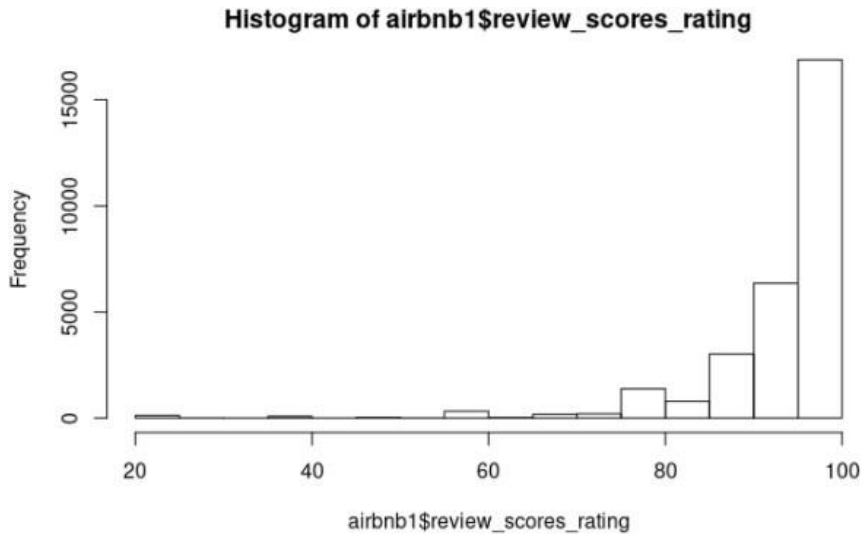
VARIANCE AND STANDARD DEVIATION

	Variance <dbl>
Price	7.061853e+04
Review Scores Rating	2.297544e-01

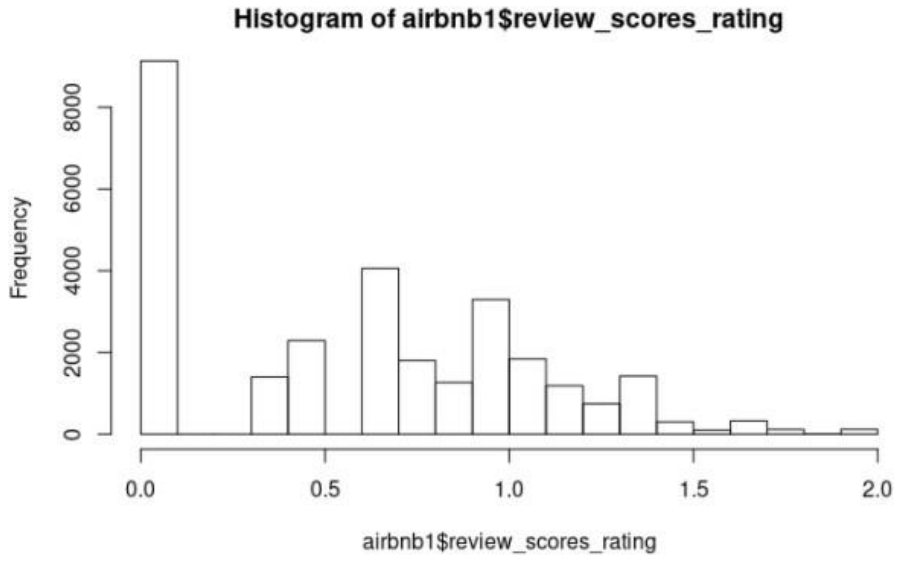
	Standard.deviation <dbl>
Price	265.741472
Review Scores Rating	0.479327

GRAPHS

Checking Skewness



Since the review rating is negatively skewed, we did the log transformation.

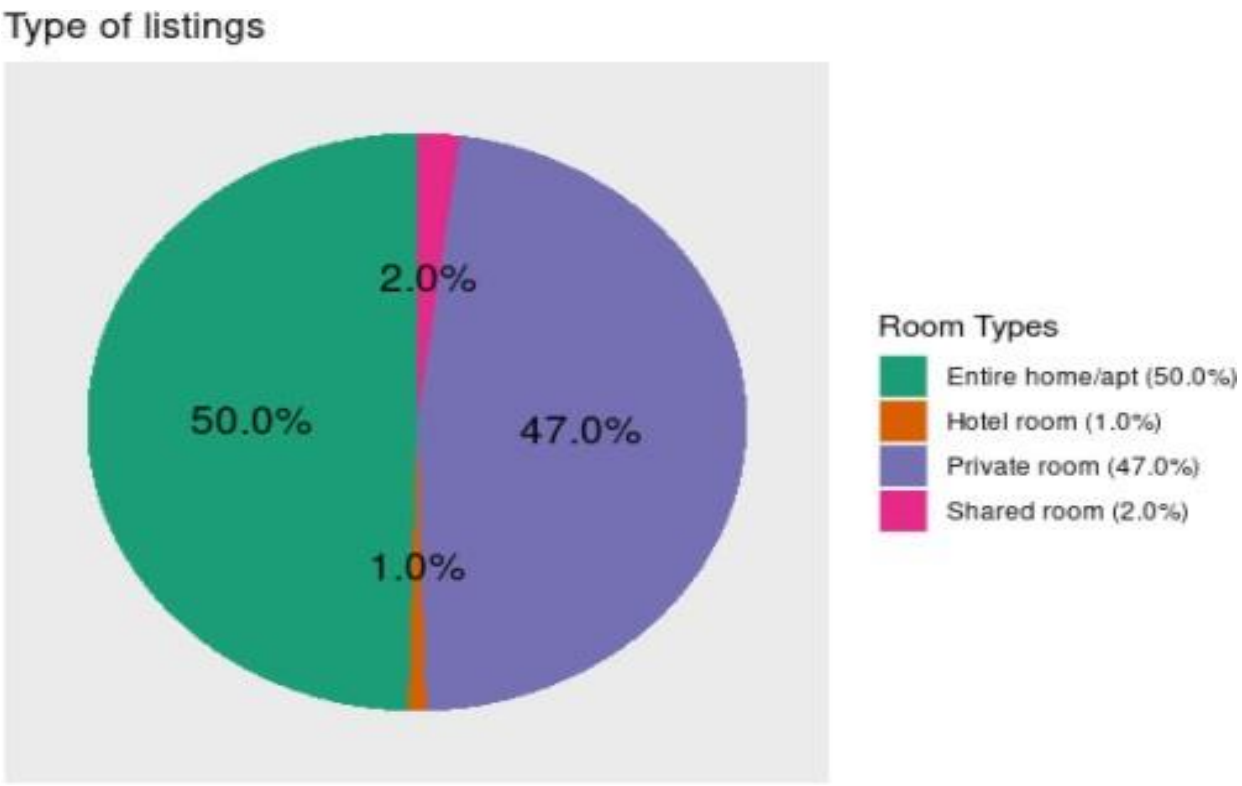


Entire home/apt is a common listing in Brooklyn, Manhattan, and Staten Island whereas Private room in Bronx and Queens. Shared room is the least common in all neighborhoods.

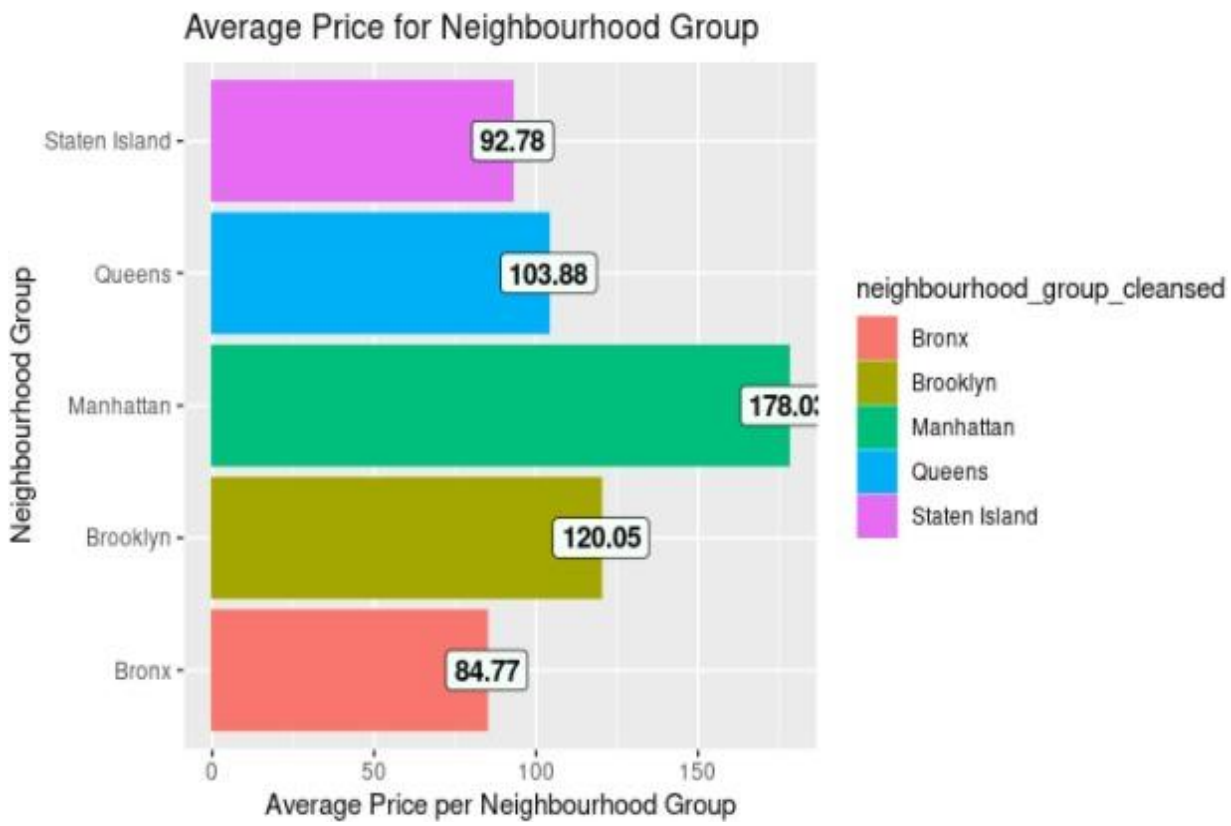
	neighbourhood_group_cleansed <chr>	Average_price_per_neighborhoodgroup <dbl>
5	Bronx	84.77359
4	Staten Island	92.78004
3	Queens	103.88463
2	Brooklyn	120.04599
1	Manhattan	178.03112

room_type <chr>	average_price <dbl>	Percent <dbl>
Entire home/apt	193.60979	27.77985
Hotel room	309.70567	44.43772
Private room	82.80253	11.88081
Shared room	110.82526	15.90162

;+

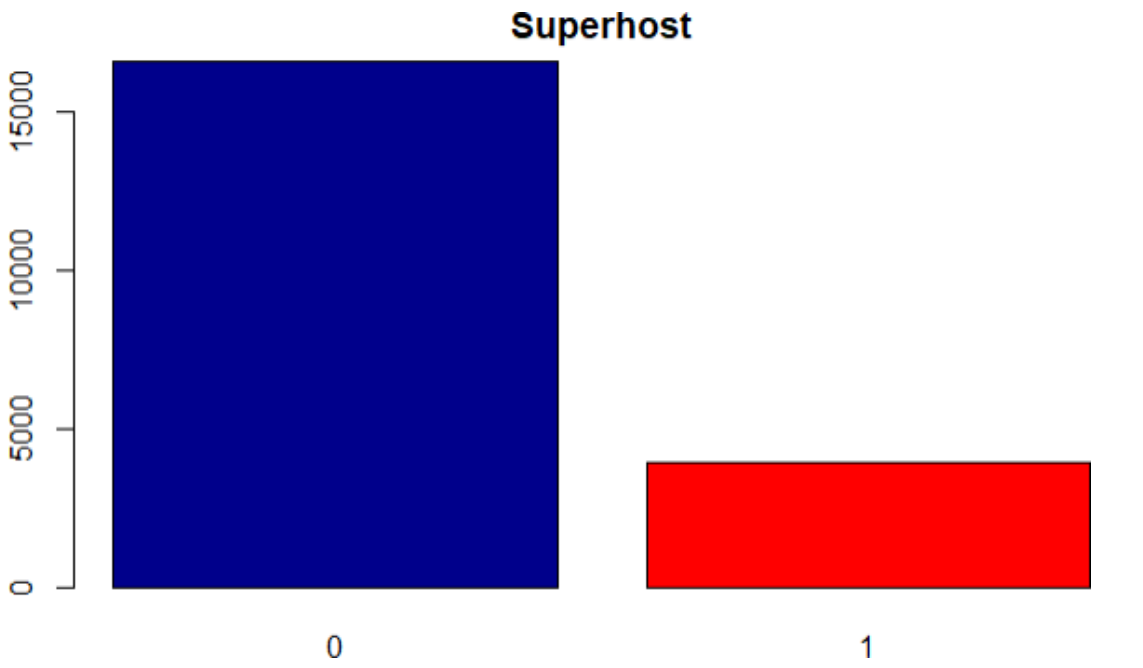


Most of the listings are entire homes or private rooms, while the shared room and hotel rooms do not represent more than 3% of the entire listings.



Average price of listings is the highest for Manhattan (127.47 USD) followed by Brookyln (108.35). One possible reason for high average price in Manhattan could be that whole apartments/home are the most common type of listings there.

Bronx has the cheapest listings with an average price of 77.69 USD.



There are approximately 4500 superhosts and approximately 15,000 hosts who are not superhosts in the entire dataset.

ANOVA: Review scores

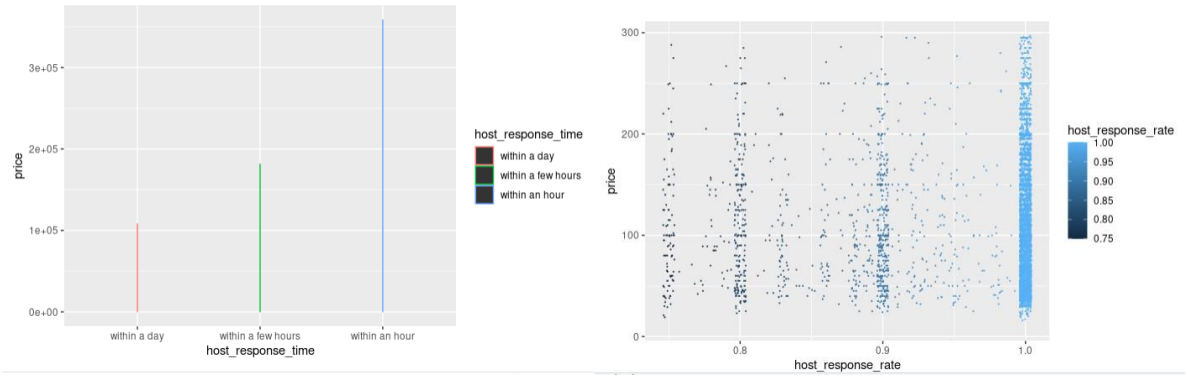
```
               Df Sum Sq Mean Sq F value   Pr(>F)
price           1      4    4.37   19.03 1.29e-05 ***
Residuals   29382   6747    0.23
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                 Df Sum Sq Mean Sq F value Pr(>F)
minimum_nights    1      0 0.00435   0.019  0.891
Residuals     29382   6751 0.22976
                 Df Sum Sq Mean Sq F value Pr(>F)
maximum_nights    1      0  0.1458   0.635  0.426
Residuals     29382   6751  0.2298
                       Df Sum Sq Mean Sq F value Pr(>F)
review_scores_accuracy  1   1938  1937.7   11828 <2e-16 ***
Residuals           29382   4813     0.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                          Df Sum Sq Mean Sq F value Pr(>F)
review_scores_cleanliness  1   2205  2204.9   14251 <2e-16 ***
Residuals              29382   4546     0.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                        Df Sum Sq Mean Sq F value Pr(>F)
review_scores_checkin    1   1118  1118.1    5832 <2e-16 ***
Residuals            29382   5633     0.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                           Df Sum Sq Mean Sq F value Pr(>F)
review_scores_communication  1   1283  1283.3    6896 <2e-16 ***
Residuals                29382   5468     0.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                        Df Sum Sq Mean Sq F value Pr(>F)
review_scores_location   1    927   927.0    4677 <2e-16 ***
Residuals            29382   5824     0.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                       Df Sum Sq Mean Sq F value Pr(>F)
review_scores_value     1   2306  2306.3   15247 <2e-16 ***
Residuals           29382   4445     0.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
               Df Sum Sq Mean Sq F value   Pr(>F)
accommodates    1      5   5.251   22.87 1.74e-06 ***
Residuals   29382   6746   0.230
---
```

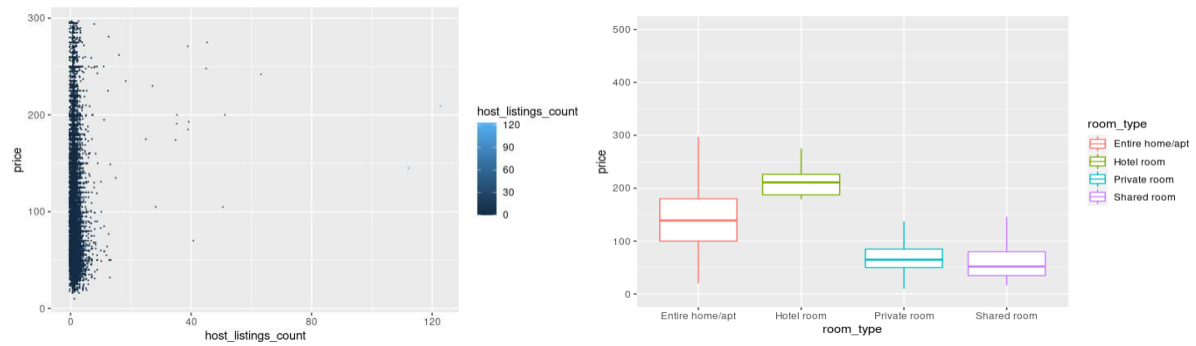
```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
              Df Sum Sq Mean Sq F value   Pr(>F)
beds           1      7   7.038   30.66 3.09e-08 ***
Residuals   29382   6744   0.230
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
              Df Sum Sq Mean Sq F value Pr(>F)
bedrooms       1      1  1.0324   4.494  0.034 *
Residuals   29382   6750  0.2297
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                     Df Sum Sq Mean Sq F value Pr(>F)
host_response_time     4     60  14.909   65.46 <2e-16 ***
Residuals          29379   6691   0.228
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                     Df Sum Sq Mean Sq F value   Pr(>F)
host_response_rate     1   12.6  12.611   61.72 4.22e-15 ***
Residuals          14966 3057.9   0.204
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
14416 observations deleted due to missingness
                       Df Sum Sq Mean Sq F value Pr(>F)
host_acceptance_rate     1     22  21.979   106.9 <2e-16 ***
Residuals            20169   4147   0.206
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
9213 observations deleted due to missingness
                      Df Sum Sq Mean Sq F value Pr(>F)
host_listings_count     1      1  0.5860   2.551   0.11
Residuals           29382   6750  0.2297
                     Df Sum Sq Mean Sq F value Pr(>F)
host_is_superhost     1    129  128.59   570.5 <2e-16 ***
Residuals         29382   6622    0.23
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                       Df Sum Sq Mean Sq F value Pr(>F)
neighbourhood_cleansed   215    168  0.7800   3.456 <2e-16 ***
Residuals              29168   6583  0.2257
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                            Df Sum Sq Mean Sq F value Pr(>F)
neighbourhood_group_cleansed     4     20   4.943   21.57 <2e-16 ***
Residuals                    29379   6731   0.229


Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                     Df Sum Sq Mean Sq F value Pr(>F)
number_of_reviews     1    132  131.80     585 <2e-16 ***
Residuals         29382   6619    0.23
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                                Df Sum Sq Mean Sq F value   Pr(>F)
calculated_host_listings_count     1     10   9.964   43.43 4.46e-11 ***
Residuals                      29382   6741   0.229
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

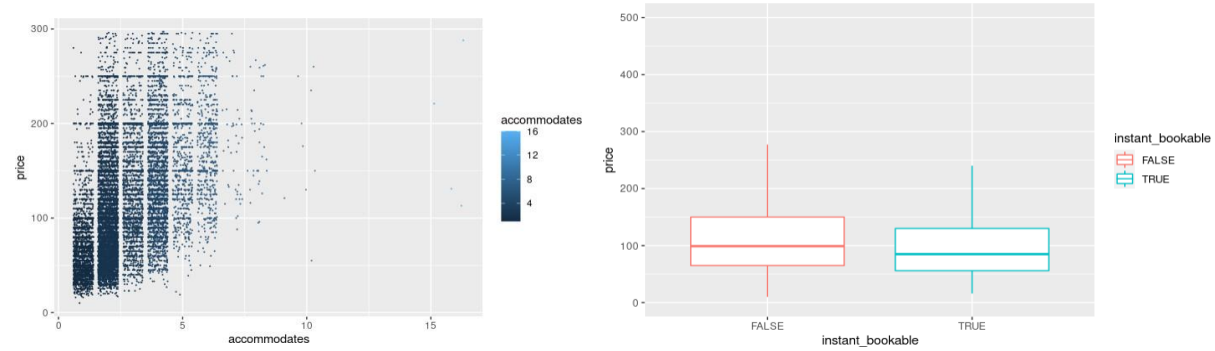
ANOVA: Superhost

```
               Df Sum Sq Mean Sq F value Pr(>F)
price           1    0.2  0.1737   1.127  0.289
Residuals   20487 3159.3  0.1542
                Df Sum Sq Mean Sq F value Pr(>F)
minimum_nights   1    0.1  0.1397   0.906  0.341
Residuals    20487 3159.4  0.1542
                Df Sum Sq Mean Sq F value Pr(>F)
maximum_nights   1    0.1 0.07693   0.499   0.48
Residuals    20487 3159.4 0.15422
                          Df Sum Sq Mean Sq F value Pr(>F)
review_scores_accuracy     1   94.9   94.90   634.4 <2e-16 ***
Residuals              20487 3064.6    0.15
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                             Df Sum Sq Mean Sq F value Pr(>F)
review_scores_cleanliness     1  140.5  140.49   953.4 <2e-16 ***
Residuals                 20487 3019.0    0.15
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                           Df Sum Sq Mean Sq F value Pr(>F)
review_scores_checkin       1   66.4   66.35   439.5 <2e-16 ***
Residuals               20487 3093.2    0.15
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                              Df Sum Sq Mean Sq F value Pr(>F)
review_scores_communication    1   63.9   63.87   422.7 <2e-16 ***
Residuals                  20487 3095.6    0.15
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                           Df Sum Sq Mean Sq F value Pr(>F)
review_scores_location      1   30.5  30.453   199.4 <2e-16 ***
Residuals               20487 3129.1   0.153
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                         Df Sum Sq Mean Sq F value Pr(>F)
review_scores_value       1     78   77.99   518.5 <2e-16 ***
Residuals             20487   3082    0.15
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                Df Sum Sq Mean Sq F value   Pr(>F)
accommodates     1      6   6.043   39.26 3.78e-10 ***
Residuals    20487   3154   0.154
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
               Df Sum Sq Mean Sq F value Pr(>F)
beds            1   12.7  12.705   82.72 <2e-16 ***
Residuals   20487 3146.8   0.154
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
               Df Sum Sq Mean Sq F value   Pr(>F)
bedrooms        1    3.6   3.645   23.66 1.16e-06 ***
Residuals   20487 3155.9   0.154
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                     Df Sum Sq Mean Sq F value Pr(>F)
host_response_time    4  417.5  104.37   779.7 <2e-16 ***
Residuals         20484 2742.0    0.13
---
```

```
                     Df Sum Sq Mean Sq F value Pr(>F)
host_response_rate    1  108.3  108.29   509.8 <2e-16 ***
Residuals          7885 1674.9    0.21
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
12602 observations deleted due to missingness
                       Df Sum Sq Mean Sq F value Pr(>F)
host_acceptance_rate    1  140.9   140.9     697 <2e-16 ***
Residuals           12092 2444.1     0.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
8395 observations deleted due to missingness
                      Df Sum Sq Mean Sq F value Pr(>F)
host_listings_count    1   23.6  23.630   154.4 <2e-16 ***
Residuals          20487 3135.9   0.153
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                      Df Sum Sq Mean Sq F value Pr(>F)
review_scores_rating   1   42.4   42.44     279 <2e-16 ***
Residuals          20487 3117.1    0.15
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                        Df Sum Sq Mean Sq F value Pr(>F)
neighbourhood_cleansed   208  133.2  0.6406   4.293 <2e-16 ***
Residuals              20280 3026.3  0.1492
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                             Df Sum Sq Mean Sq F value Pr(>F)
neighbourhood_group_cleansed    4   34.5   8.618   56.49 <2e-16 ***
Residuals                   20484 3125.0   0.153
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                     Df Sum Sq Mean Sq F value Pr(>F)
number_of_reviews     1  395.4   395.4    2931 <2e-16 ***
Residuals         20487 2764.1     0.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                               Df Sum Sq Mean Sq F value Pr(>F)
calculated_host_listings_count     1  111.7  111.69   750.8 <2e-16 ***
Residuals                      20487 3047.8    0.15
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```







The plots show no correlations between the price and the majority of the parameters: the host response time, the host response rate, the host total listings count, the number of accommodates, or if the listing is instant bookable or not. There is a possible correlation between the price and room type - the hotel homes are generally more expensive than the other types, and the shared rooms are the cheapest ones.

CORRELATION:

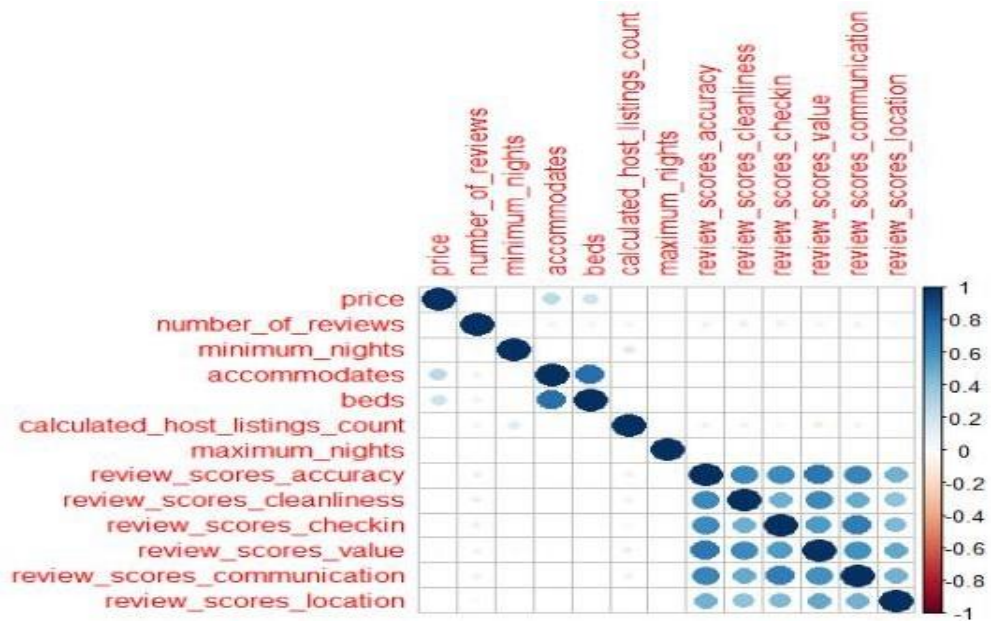
	price	number_of_reviews	minimum_nights
price	1.000000000	-0.030529964	0.0142905749
number_of_reviews	-0.030529964	1.000000000	-0.0456490591
minimum_nights	0.014290575	-0.045649059	1.0000000000
accommodates	0.267196963	0.062990667	0.0052830928
beds	0.209436914	0.060463321	0.0145776455
calculated_host_listings_count	-0.010242357	-0.065246223	0.1358333903
maximum_nights	0.000651205	-0.004827592	0.0032070230
review_scores_accuracy	-0.006288536	0.085346613	-0.0189712750
review_scores_cleanliness	0.027707788	0.098056626	-0.0326006610
review_scores_checkin	-0.021453439	0.087193493	-0.0109017608
review_scores_value	-0.033777569	0.066746150	-0.0317212777
review_scores_communication	-0.009663760	0.074530923	-0.0257863225
review_scores_location	0.028378668	0.045041878	-0.0009917445

	accommodates	beds
price	0.267196963	0.2094369138
number_of_reviews	0.062990667	0.0604633215
minimum_nights	0.005283093	0.0145776455
accommodates	1.000000000	0.7548561276
beds	0.754856128	1.0000000000
calculated_host_listings_count	-0.027395501	-0.0251590411
maximum_nights	-0.003731813	-0.0040569021
review_scores_accuracy	-0.013410589	-0.0165627945
review_scores_cleanliness	0.032927589	0.0203680208
review_scores_checkin	-0.004593946	-0.0004656877
review_scores_value	-0.030628557	-0.0225930660
review_scores_communication	-0.012986566	-0.0114274922
review_scores_location	-0.010638241	-0.0131754768

	calculated_host_listings_count	maximum_nights
price	-0.010242357	0.000651205
number_of_reviews	-0.065246223	-0.004827592
minimum_nights	0.135833390	0.003207023
accommodates	-0.027395501	-0.003731813
beds	-0.025159041	-0.004056902
calculated_host_listings_count	1.000000000	-0.001428853
maximum_nights	-0.001428853	1.000000000
review_scores_accuracy	-0.080097652	-0.001098610
review_scores_cleanliness	-0.065002687	0.001442238
review_scores_checkin	-0.051536012	-0.007703355
review_scores_value	-0.109517913	-0.016186184

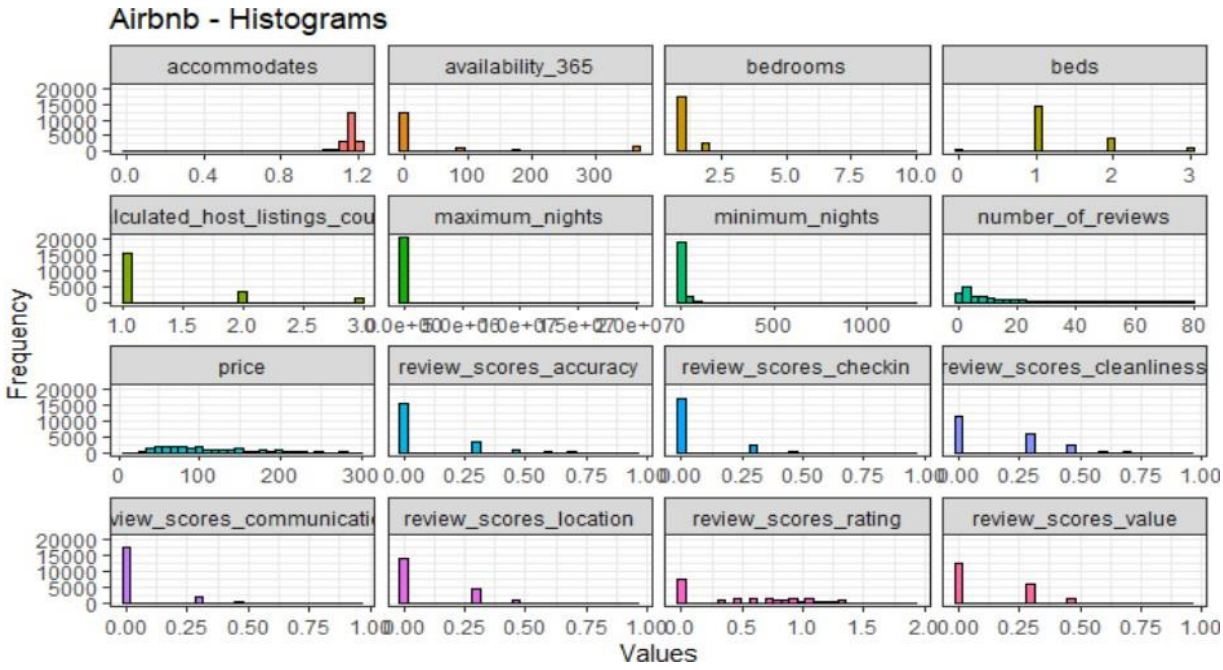
```
review_scores_communication                    -0.076433206    -0.017693818
review_scores_location                         -0.017487173    -0.006036953
                                 review_scores_accuracy review_scores_cleanliness
price                                      -0.006288536                0.027707788
number_of_reviews                           0.085346613                0.098056626
minimum_nights                             -0.018971275               -0.032600661
accommodates                               -0.013410589                0.032927589
beds                                       -0.016562794                0.020368021
calculated_host_listings_count             -0.080097652               -0.065002687
maximum_nights                             -0.001098610                0.001442238
review_scores_accuracy                      1.000000000                0.641600625
review_scores_cleanliness                   0.641600625                1.000000000
review_scores_checkin                       0.624665129                0.483162893
review_scores_value                         0.720776906                0.638105888
review_scores_communication                 0.651085179                0.509209291
review_scores_location                      0.478515248                0.400352850
                                 review_scores_checkin review_scores_value
price                                      -0.0214534395          -0.03377757
number_of_reviews                           0.0871934931           0.06674615
minimum_nights                             -0.0109017608          -0.03172128
accommodates                               -0.0045939465          -0.03062856
beds                                       -0.0004656877          -0.02259307
calculated_host_listings_count             -0.0515360116          -0.10951791
maximum_nights                             -0.0077033551          -0.01618618
review_scores_accuracy                      0.6246651290           0.72077691
review_scores_cleanliness                   0.4831628925           0.63810589
review_scores_checkin                       1.0000000000           0.56349424
review_scores_value                         0.5634942448           1.00000000
review_scores_communication                 0.6982832775           0.60839388
review_scores_location                      0.4421456329           0.51464193
                                 review_scores_communication review_scores_location
price                                            -0.00966376             0.0283786676
number_of_reviews                                 0.07453092             0.0450418782
minimum_nights                                   -0.02578632            -0.0009917445
accommodates                                     -0.01298657            -0.0106382408
beds                                             -0.01142749            -0.0131754768
calculated_host_listings_count                   -0.07643321            -0.0174871730
maximum_nights                                   -0.01769382            -0.0060369534
review_scores_accuracy                            0.65108518             0.4785152479
review_scores_cleanliness                         0.50920929             0.4003528498
```

The correlation is positive when two variables tend to move in the same direction. Good correlation is considered to be more than 0.6. In our model, beds and accommodates; review scores cleanliness, value and communication are highly correlated to review scores checkin.

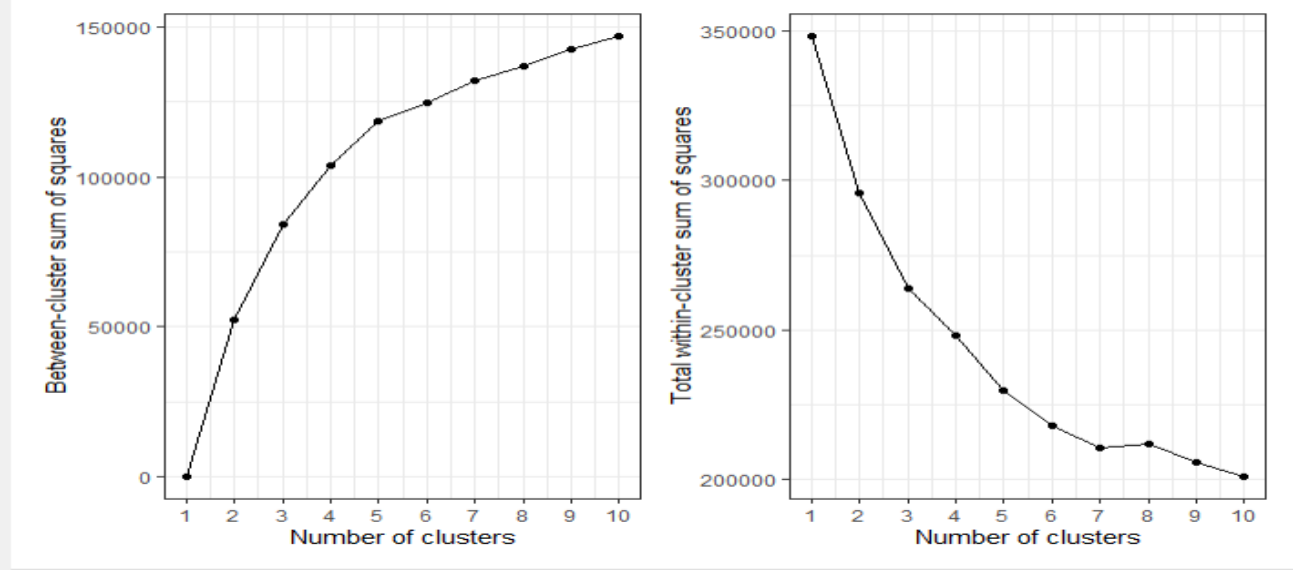


Predictive Analytics

CLUSTERING ANALYSIS: Superhost




```
  accommodates  bedrooms      beds      price minimum_nights maximum_nights
1   -1.4128217  1.698628  1.563030  0.9795410     0.10660651    0.018103916
2    0.3043039 -0.365863 -0.336657 -0.2109807    -0.02296169   -0.003899355
  availability_365 number_of_reviews review_scores_rating review_scores_accuracy
1       0.17581982        0.11548390         -0.009295260           -0.010812220
2      -0.03786937       -0.02487377          0.002002081            0.002328815
  review_scores_cleanliness review_scores_checkin review_scores_communication
1               -0.06862144          -0.025323938                  -0.01280470
2                0.01478019           0.005454456                   0.00275797
  review_scores_location review_scores_value calculated_host_listings_count
1            0.005644534         0.032750536                    -0.017151248
2           -0.001215761        -0.007054051                     0.003694162
    superhost
1  0.14399554
2 -0.03101482
[1]  3631 16858
[1] 36927.52
[1]  79781.64 231586.84
[1] 311368.5
[1] 348296
```



The optimal number of clusters is 5.

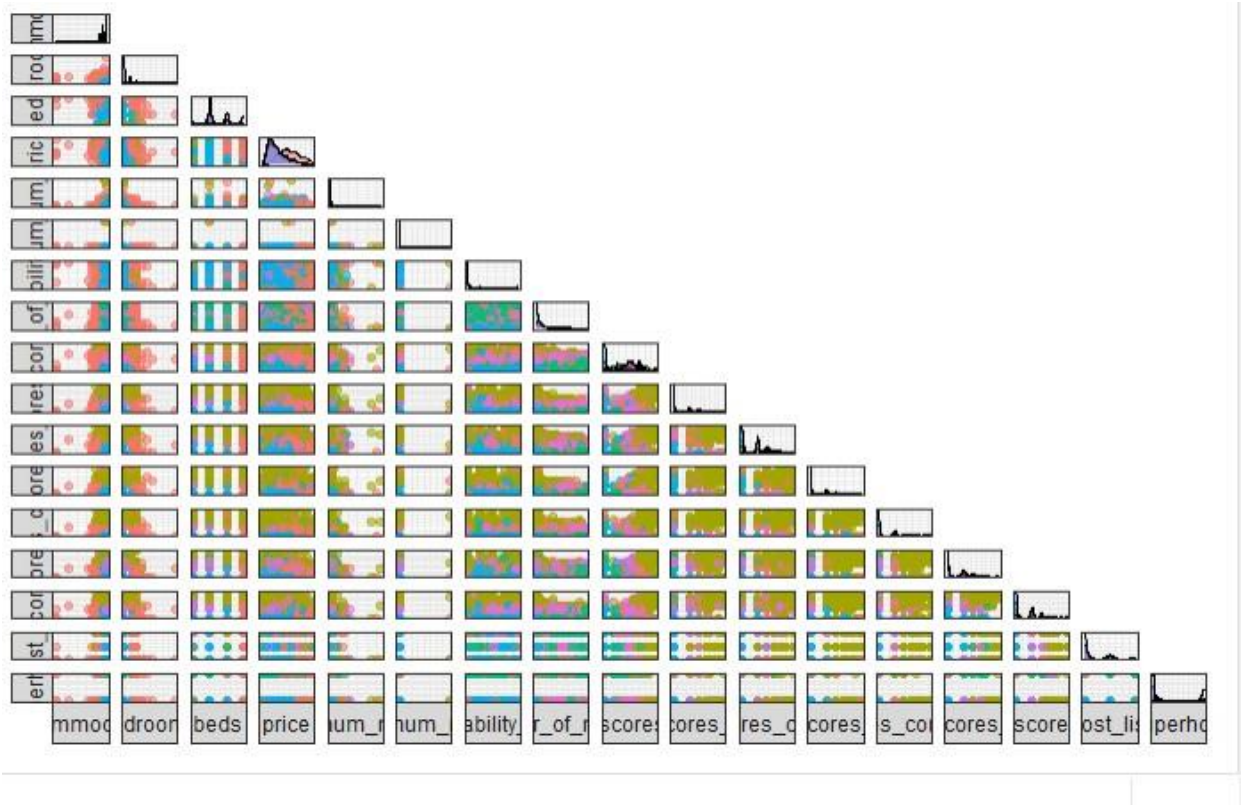
Group.1 <int>	accommodates <dbl>	bedrooms <dbl>	beds <dbl>	price <dbl>	minimum_nights <dbl>	maximum_nights <dbl>	availability_365 <dbl>	number_of_reviews <dbl>	review_scores_rating <dbl>
1	-1.56875215	2.0763330	1.66944417	1.04448586	0.110928955	0.027347342	0.16367006	0.0954503	-0.1306883
2	0.09909293	-0.1031751	-0.08327251	-0.24008455	0.049045229	0.044056583	0.14094783	-0.2605934	1.4623641
3	0.23273980	-0.3436507	-0.19913035	-0.18855935	-0.002654186	-0.010177886	0.34447196	0.8239961	-0.2538268
4	0.27979125	-0.3453063	-0.32857876	-0.07998313	-0.026989812	-0.009897596	-0.25043490	-0.4121862	-0.7135471
5	0.23948150	-0.3426828	-0.24803059	-0.23519163	-0.036507073	-0.009846586	0.01336688	0.1349865	0.7893092

5 rows | 1-10 of 18 columns

review_scores_accuracy <dbl>	review_scores_cleanliness <dbl>	review_scores_checkin <dbl>	review_scores_communication <dbl>	review_scores_location <dbl>	review_scores_value <dbl>
-0.1659704	-0.1837006	-0.17403188	-0.17633332	-0.07664586	-0.1087682
1.9146417	1.4019682	2.09004135	2.23572147	1.20661716	1.5793879
-0.4046310	-0.4759512	-0.33051969	-0.32840813	-0.23635690	-0.3925580
-0.4430522	-0.4846573	-0.28635535	-0.30972513	-0.35031948	-0.5701333
0.3162937	0.6352565	-0.03956347	-0.05932025	0.28178904	0.6025511

5 rows | 11-16 of 18 columns

calculated_host_listings_count <dbl>	superhost <dbl>
-0.01825280	0.1101836
0.16828819	-0.4150232
0.42447920	1.9634226
-0.24385555	-0.4850850
0.03969047	-0.4644029


mmod
droom
beds
price
um_r
num_
ability_
r_of_r
scores
cores_
res_c
cores_
s_cor
cores_
score
ost_lis
perho

REGRESSION ANALYSIS

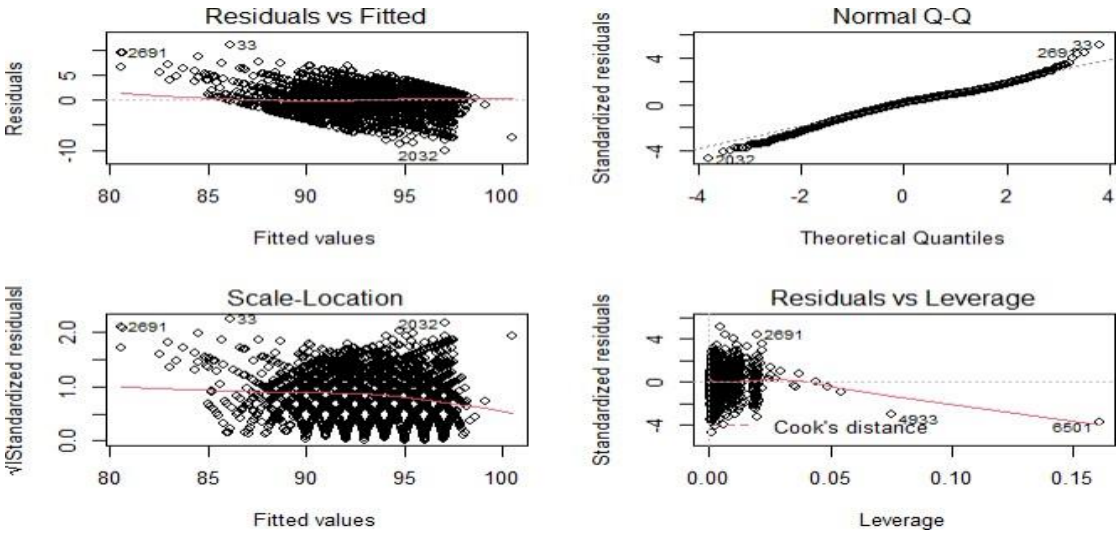
```
Call:
lm(formula = review_scores_rating ~ neighbourhood_group_cleansed +
    room_type + number_of_reviews + price + calculated_host_listings_count +
    review_scores_accuracy + review_scores_cleanliness + review_scores_checkin +
    review_scores_communication + review_scores_location + review_scores_value +
    availability_365, data = airbnb_train)

Residuals:
     Min       1Q   Median       3Q      Max
-10.0895  -1.2529   0.2924   1.4962  10.8952

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                       25.1440680  0.8492475  29.607  < 2e-16 ***
neighbourhood_group_cleansedBrooklyn  0.0530867  0.1342479   0.395  0.69253
neighbourhood_group_cleansedManhattan -0.0310434  0.1383344  -0.224  0.82245
neighbourhood_group_cleansedQueens   -0.1069939  0.1426465  -0.750  0.45324
neighbourhood_group_cleansedStaten Island  0.0171387  0.2467930   0.069  0.94464
room_typeHotel room               -0.1258103  0.2932338  -0.429  0.66790
room_typePrivate room             -0.1662348  0.0573528  -2.898  0.00376 **
room_typeShared room              -0.3490306  0.1879039  -1.857  0.06328 .
number_of_reviews                 -0.0012413  0.0003784  -3.281  0.00104 **
price                              0.0007896  0.0001758   4.493 7.15e-06 ***
calculated_host_listings_count    -0.0033994  0.0019860  -1.712  0.08700 .
review_scores_accuracy             1.8315867  0.0670910  27.300  < 2e-16 ***
review_scores_cleanliness          1.5931901  0.0469934  33.902  < 2e-16 ***
review_scores_checkin              0.8364936  0.0803774  10.407  < 2e-16 ***
review_scores_communication        1.1014072  0.0806409  13.658  < 2e-16 ***
review_scores_location             0.4902815  0.0529103   9.266  < 2e-16 ***
review_scores_value                1.3685248  0.0581575  23.531  < 2e-16 ***
availability_365                  -0.0004071  0.0001882  -2.163  0.03057 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.15 on 6798 degrees of freedom
Multiple R-squared:  0.6076,    Adjusted R-squared:  0.6066
F-statistic: 619.2 on 17 and 6798 DF,  p-value: < 2.2e-16
```

Model Observations:

MSE: 2.15 on 6798 degrees of freedom

R-squared: 0.6076

Adjusted R-squared: 0.6066

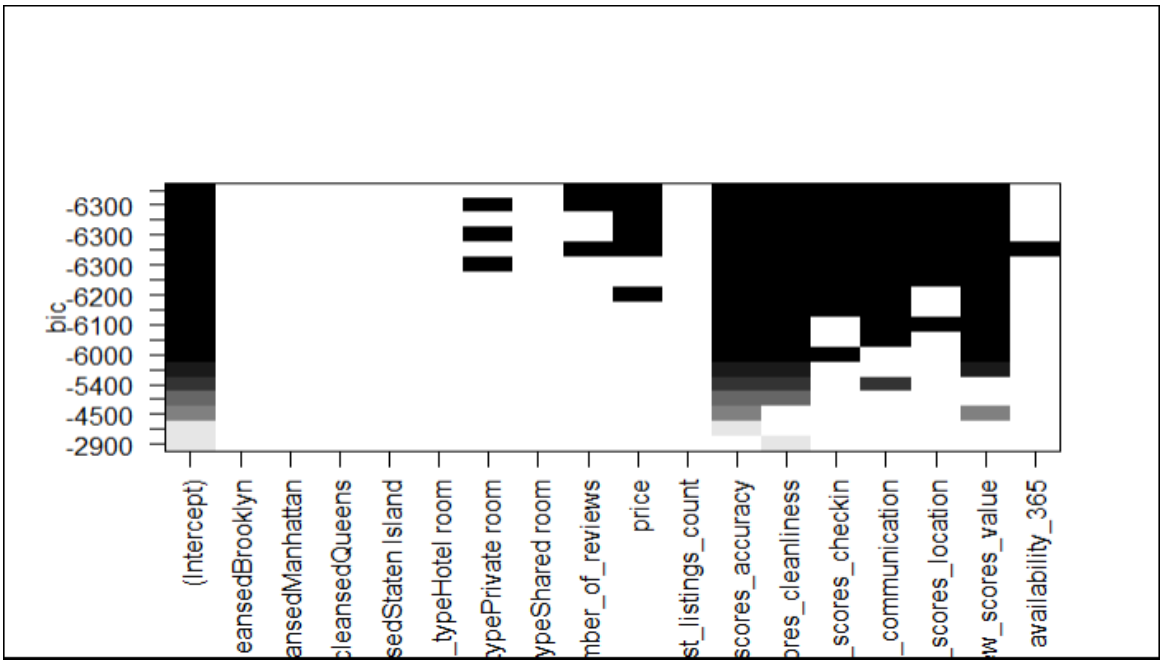



```
      review_scores_location review_scores_value availability_365
1  ( 1 ) " "                 " "                 " "
1  ( 2 ) " "                 " "                 " "
2  ( 1 ) " "                 " "                 " "
2  ( 2 ) " "                 "*"                 " "
3  ( 1 ) " "                 "*"                 " "
3  ( 2 ) " "                 " "                 " "
4  ( 1 ) " "                 "*"                 " "
4  ( 2 ) " "                 "*"                 " "
5  ( 1 ) " "                 "*"                 " "
5  ( 2 ) "*"                 "*"                 " "
6  ( 1 ) "*"                 "*"                 " "
6  ( 2 ) " "                 "*"                 " "
7  ( 1 ) "*"                 "*"                 " "
7  ( 2 ) "*"                 "*"                 " "
8  ( 1 ) "*"                 "*"                 " "
8  ( 2 ) "*"                 "*"                 " "
9  ( 1 ) "*"                 "*"                 " "
9  ( 2 ) "*"                 "*"                 "*"
```

```
      room_typePrivate room room_typeShared room number_of_reviews price calculated_host_listings_count
1  ( 1 ) " "                 " "                 " "               " "   " "
1  ( 2 ) " "                 " "                 " "               " "   " "
2  ( 1 ) " "                 " "                 " "               " "   " "
2  ( 2 ) " "                 " "                 " "               " "   " "
3  ( 1 ) " "                 " "                 " "               " "   " "
3  ( 2 ) " "                 " "                 " "               " "   " "
4  ( 1 ) " "                 " "                 " "               " "   " "
4  ( 2 ) " "                 " "                 " "               " "   " "
5  ( 1 ) " "                 " "                 " "               " "   " "
5  ( 2 ) " "                 " "                 " "               " "   " "
6  ( 1 ) " "                 " "                 " "               " "   " "
6  ( 2 ) " "                 " "                 " "               "*"   " "
7  ( 1 ) " "                 " "                 " "               "*"   " "
7  ( 2 ) "*"                 " "                 " "               "*"   " "
8  ( 1 ) "*"                 " "                 "*"               "*"   " "
8  ( 2 ) "*"                 " "                 "*"               "*"   " "
9  ( 1 ) "*"                 " "                 "*"               "*"   " "
9  ( 2 ) "*"                 " "                 "*"               "*"   " "
```

```
      review_scores_accuracy review_scores_cleanliness review_scores_checkin review_scores_communication
1  ( 1 ) "*"                 " "                       " "                 " "
1  ( 2 ) " "                 "*"                       " "                 " "
2  ( 1 ) "*"                 "*"                       " "                 " "
2  ( 2 ) "*"                 " "                       " "                 " "
3  ( 1 ) "*"                 "*"                       " "                 " "
3  ( 2 ) "*"                 "*"                       " "                 "*"
4  ( 1 ) "*"                 "*"                       " "                 "*"
4  ( 2 ) "*"                 "*"                       "*"                 "*"
5  ( 1 ) "*"                 "*"                       "*"                 "*"
5  ( 2 ) "*"                 "*"                       "*"                 "*"
6  ( 1 ) "*"                 "*"                       "*"                 "*"
6  ( 2 ) "*"                 "*"                       "*"                 "*"
7  ( 1 ) "*"                 "*"                       "*"                 "*"
7  ( 2 ) "*"                 "*"                       "*"                 "*"
8  ( 1 ) "*"                 "*"                       "*"                 "*"
8  ( 2 ) "*"                 "*"                       "*"                 "*"
9  ( 1 ) "*"                 "*"                       "*"                 "*"
9  ( 2 ) "*"                 "*"                       "*"                 "*"
```

```
Subset selection object
Call: regsubsets.formula(review_scores_rating ~ neighbourhood_group_cleansed +
    room_type + number_of_reviews + price + calculated_host_listings_count +
    review_scores_accuracy + review_scores_cleanliness + review_scores_checkin +
    review_scores_communication + review_scores_location + review_scores_value +
    availability_365, data = airbnb_train, nbest = 2, nvmax = 9)
17 Variables  (and intercept)
                                      Forced in Forced out
neighbourhood_group_cleansedBrooklyn      FALSE      FALSE
neighbourhood_group_cleansedManhattan     FALSE      FALSE
neighbourhood_group_cleansedQueens        FALSE      FALSE
neighbourhood_group_cleansedStaten Island FALSE      FALSE
room_typeHotel room                       FALSE      FALSE
room_typePrivate room                     FALSE      FALSE
room_typeShared room                      FALSE      FALSE
number_of_reviews                         FALSE      FALSE
price                                     FALSE      FALSE
calculated_host_listings_count            FALSE      FALSE
review_scores_accuracy                    FALSE      FALSE
review_scores_cleanliness                 FALSE      FALSE
review_scores_checkin                     FALSE      FALSE
review_scores_communication               FALSE      FALSE
review_scores_location                    FALSE      FALSE
review_scores_value                       FALSE      FALSE
availability_365                          FALSE      FALSE
2 subsets of each size up to 9
Selection Algorithm: exhaustive
         neighbourhood_group_cleansedBrooklyn neighbourhood_group_cleansedManhattan
1  ( 1 ) " "                                  " "
1  ( 2 ) " "                                  " "
2  ( 1 ) " "                                  " "
2  ( 2 ) " "                                  " "
3  ( 1 ) " "                                  " "
3  ( 2 ) " "                                  " "
4  ( 1 ) " "                                  " "
4  ( 2 ) " "                                  " "
5  ( 1 ) " "                                  " "
5  ( 2 ) " "                                  " "
6  ( 1 ) " "                                  " "
6  ( 2 ) " "                                  " "
7  ( 1 ) " "                                  " "
7  ( 2 ) " "                                  " "
8  ( 1 ) " "                                  " "
8  ( 2 ) " "                                  " "
9  ( 1 ) " "                                  " "
9  ( 2 ) " "                                  " "
         neighbourhood_group_cleansedQueens neighbourhood_group_cleansedStaten Island room_typeHotel room
1  ( 1 ) " "                                " "                                       " "
1  ( 2 ) " "                                " "                                       " "
2  ( 1 ) " "                                " "                                       " "
2  ( 2 ) " "                                " "                                       " "
3  ( 1 ) " "                                " "                                       " "
3  ( 2 ) " "                                " "                                       " "
```



```
Call:
lm(formula = review_scores_rating ~ room_type + number_of_reviews +
    price + review_scores_accuracy + review_scores_cleanliness +
    review_scores_checkin + review_scores_communication + review_scores_location +
    review_scores_value, data = airbnb_train, subset = availability_365,
    nbest = 2, nvmax = 9)

Residuals:
    Min      1Q  Median      3Q     Max
-8.0377 -0.7490  0.2984  1.1523 13.0339

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 21.1157940  0.8378415  25.203  < 2e-16 ***
room_typeHotel room          5.5557543  0.3724456  14.917  < 2e-16 ***
room_typePrivate room       -0.1061646  0.0613394  -1.731   0.0835 .
room_typeShared room        -5.0172113  0.1603455 -31.290  < 2e-16 ***
number_of_reviews           -0.0048777  0.0004175 -11.683  < 2e-16 ***
price                        0.0017017  0.0001413  12.040  < 2e-16 ***
review_scores_accuracy       3.1033903  0.0760864  40.788  < 2e-16 ***
review_scores_cleanliness    2.1124615  0.0519566  40.658  < 2e-16 ***
review_scores_checkin        1.4755657  0.0812286  18.166  < 2e-16 ***
review_scores_communication -0.4314776  0.0936710  -4.606 4.19e-06 ***
review_scores_location       0.0459386  0.0506684   0.907   0.3646
review_scores_value          1.3549523  0.0586379  23.107  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.935 on 5680 degrees of freedom
Multiple R-squared:  0.7222,    Adjusted R-squared:  0.7217
F-statistic:  1343 on 11 and 5680 DF,  p-value: < 2.2e-16
```

Model Observations:

MSE: 1.935 on 5680 degrees of freedom

R-squared: 0.7222

Adjusted R-squared: 0.7217

Stepwise Selection using AIC (Direction = " both")

```
Start:  AIC=16792.69
review_scores_rating ~ 1

                            Df Sum of Sq   RSS   AIC
+ review_scores_accuracy     1   30505.6 49548 13525
+ review_scores_cleanliness  1   28106.3 51947 13847
+ review_scores_value        1   27262.9 52791 13957
+ review_scores_communication  1   17929.3 62124 15066
+ review_scores_checkin      1   15746.7 64307 15302
+ review_scores_location     1    7247.0 72806 16148
+ room_type                  3    1593.6 78460 16662
+ number_of_reviews          1     475.6 79578 16754
+ price                      1     304.0 79750 16769
<none>                                  80054 16793

Step:  AIC=13524.68
review_scores_rating ~ review_scores_accuracy

                            Df Sum of Sq   RSS   AIC
+ review_scores_cleanliness  1   10528.5 39019 11898
+ review_scores_value        1    8334.6 41213 12271
+ review_scores_communication  1    4548.1 45000 12870
+ review_scores_checkin      1    3681.1 45867 13000
+ review_scores_location     1    1931.7 47616 13256
+ room_type                  3     338.4 49210 13484
+ price                      1     309.3 49239 13484
<none>                                  49548 13525
+ number_of_reviews          1       7.8 49540 13526
- review_scores_accuracy     1   30505.6 80054 16793

Step:  AIC=11898.47
review_scores_rating ~ review_scores_accuracy + review_scores_cleanliness

                            Df Sum of Sq   RSS   AIC
+ review_scores_value        1    4271.7 34748 11110
+ review_scores_communication  1    2971.3 36048 11361
+ review_scores_checkin      1    2355.8 36664 11476
+ review_scores_location     1    1372.9 37647 11656
+ room_type                  3     153.7 38866 11878
+ price                      1      95.5 38924 11884
<none>                                  39019 11898
+ number_of_reviews          1       7.1 39012 11899
- review_scores_cleanliness  1   10528.5 49548 13525
- review_scores_accuracy     1   12927.8 51947 13847
```

```
Step:  AIC=11110.19
review_scores_rating ~ review_scores_accuracy + review_scores_cleanliness +
    review_scores_value

                             Df Sum of Sq   RSS   AIC
+ review_scores_communication  1    1999.5 32748 10708
+ review_scores_checkin        1    1537.1 33211 10804
+ review_scores_location       1     663.8 34084 10981
+ room_type                    3     261.4 34486 11065
+ price                        1     218.0 34530 11069
+ number_of_reviews            1      28.9 34719 11106
<none>                                    34748 11110
- review_scores_value          1    4271.7 39019 11898
- review_scores_cleanliness    1    6465.6 41213 12271
- review_scores_accuracy       1    6555.0 41303 12286

Step:  AIC=10708.24
review_scores_rating ~ review_scores_accuracy + review_scores_cleanliness +
    review_scores_value + review_scores_communication

                             Df Sum of Sq   RSS   AIC
+ review_scores_checkin        1     527.5 32221 10600
+ review_scores_location       1     492.9 32255 10607
+ price                        1     213.2 32535 10666
+ room_type                    3     178.1 32570 10677
+ number_of_reviews            1      37.4 32711 10702
<none>                                    32748 10708
- review_scores_communication  1    1999.5 34748 11110
- review_scores_value          1    3299.9 36048 11361
- review_scores_accuracy       1    4363.5 37112 11559
- review_scores_cleanliness    1    5838.2 38586 11824

Step:  AIC=10599.56
review_scores_rating ~ review_scores_accuracy + review_scores_cleanliness +
    review_scores_value + review_scores_communication + review_scores_checkin

                             Df Sum of Sq   RSS   AIC
+ review_scores_location       1     462.4 31758 10503
+ price                        1     228.4 31992 10553
+ room_type                    3     167.5 32053 10570
+ number_of_reviews            1      50.6 32170 10591
<none>                                    32221 10600
- review_scores_checkin        1     527.5 32748 10708
- review_scores_communication  1     989.9 33211 10804
- review_scores_value          1    3033.7 35254 11211
- review_scores_accuracy       1    3787.0 36008 11355
- review_scores_cleanliness    1    5661.6 37882 11701
```

```
Step:  AIC=10503.04
review_scores_rating ~ review_scores_accuracy + review_scores_cleanliness +
    review_scores_value + review_scores_communication + review_scores_checkin +
    review_scores_location

                             Df Sum of Sq   RSS   AIC
+ price                        1     170.1 31588 10468
+ room_type                    3     129.3 31629 10481
+ number_of_reviews            1      59.4 31699 10492
<none>                                    31758 10503
- review_scores_location       1     462.4 32221 10600
- review_scores_checkin        1     496.9 32255 10607
- review_scores_communication  1     904.5 32663 10692
- review_scores_value          1    2588.3 34347 11035
- review_scores_accuracy       1    3535.2 35294 11220
- review_scores_cleanliness    1    5636.1 37395 11614

Step:  AIC=10468.45
review_scores_rating ~ review_scores_accuracy + review_scores_cleanliness +
    review_scores_value + review_scores_communication + review_scores_checkin +
    review_scores_location + price

                             Df Sum of Sq   RSS   AIC
+ number_of_reviews            1      48.5 31540 10460
+ room_type                    3      59.0 31529 10462
<none>                                    31588 10468
- price                        1     170.1 31758 10503
- review_scores_location       1     404.0 31992 10553
- review_scores_checkin        1     511.5 32100 10576
- review_scores_communication  1     899.5 32488 10658
- review_scores_value          1    2686.2 34275 11023
- review_scores_accuracy       1    3548.6 35137 11192
- review_scores_cleanliness    1    5402.8 36991 11543

Step:  AIC=10459.97
review_scores_rating ~ review_scores_accuracy + review_scores_cleanliness +
    review_scores_value + review_scores_communication + review_scores_checkin +
    review_scores_location + price + number_of_reviews

                             Df Sum of Sq   RSS   AIC
+ room_type                    3      60.0 31480 10453
<none>                                    31540 10460
- number_of_reviews            1      48.5 31588 10468
- price                        1     159.1 31699 10492
- review_scores_location       1     412.9 31953 10547
- review_scores_checkin        1     523.5 32063 10570
- review_scores_communication  1     898.0 32438 10649
- review_scores_value          1    2702.0 34242 11018
- review_scores_accuracy       1    3575.2 35115 11190
- review_scores_cleanliness    1    5435.9 36976 11542
```

```
Step:  AIC=10452.99
review_scores_rating ~ review_scores_accuracy + review_scores_cleanliness +
    review_scores_value + review_scores_communication + review_scores_checkin +
    review_scores_location + price + number_of_reviews + room_type

                            Df Sum of Sq   RSS   AIC
<none>                                   31480 10453
- room_type                  3      60.0 31540 10460
- number_of_reviews          1      49.5 31529 10462
- price                      1      91.8 31572 10471
- review_scores_location     1     399.7 31880 10537
- review_scores_checkin      1     509.9 31990 10560
- review_scores_communication 1     875.9 32356 10638
- review_scores_value        1    2730.8 34211 11018
- review_scores_accuracy     1    3489.7 34970 11168
- review_scores_cleanliness  1    5323.9 36804 11516

Call:
lm(formula = review_scores_rating ~ review_scores_accuracy +
    review_scores_cleanliness + review_scores_value + review_scores_communication +
    review_scores_checkin + review_scores_location + price +
    number_of_reviews + room_type, data = airbnb_train)

Coefficients:
               (Intercept)       review_scores_accuracy    review_scores_cleanliness
                24.7536284                    1.8413129                    1.5873319
       review_scores_value  review_scores_communication       review_scores_checkin
                 1.3961638                    1.1090685                    0.8422442
    review_scores_location                        price           number_of_reviews
                 0.4776080                    0.0007713                   -0.0012254
      room_typeHotel room       room_typePrivate room        room_typeShared room
                -0.1913761                   -0.1834034                   -0.3894232
```

```
Call:
lm(formula = review_scores_rating ~ price + number_of_reviews +
    review_scores_accuracy + review_scores_checkin + review_scores_cleanliness +
    review_scores_communication + review_scores_location + review_scores_value +
    room_type, data = airbnb_train, nbest = 2, nvmax = 9)

Residuals:
     Min       1Q   Median       3Q      Max
-10.1137  -1.2489   0.3009   1.4957  11.0486

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                24.7536284  0.8340064  29.680  < 2e-16 ***
price                       0.0007713  0.0001731   4.455 8.51e-06 ***
number_of_reviews          -0.0012254  0.0003746  -3.271  0.00108 **
review_scores_accuracy      1.8413129  0.0670455  27.464  < 2e-16 ***
review_scores_checkin       0.8422442  0.0802286  10.498  < 2e-16 ***
review_scores_cleanliness   1.5873319  0.0467937  33.922  < 2e-16 ***
review_scores_communication 1.1090685  0.0806037  13.760  < 2e-16 ***
review_scores_location      0.4776080  0.0513873   9.294  < 2e-16 ***
review_scores_value         1.3961638  0.0574683  24.295  < 2e-16 ***
room_typeHotel room        -0.1913761  0.2920388  -0.655  0.51229
room_typePrivate room      -0.1834034  0.0569807  -3.219  0.00129 **
room_typeShared room       -0.3894232  0.1875915  -2.076  0.03794 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.151 on 6804 degrees of freedom
Multiple R-squared:  0.6068,    Adjusted R-squared:  0.6061
F-statistic: 954.4 on 11 and 6804 DF,  p-value: < 2.2e-16
```

Model Observations:

MSE: 2.151 on 6840 degrees of freedom

R-squared: 0.6068

Adjusted R-squared: 0.6061

```
Start:  AIC=16799.52
review_scores_rating ~ 1

                            Df Sum of Sq   RSS   AIC
+ review_scores_accuracy     1   30505.6 49548 13521
+ review_scores_cleanliness  1   28106.3 51947 13843
+ review_scores_value        1   27262.9 52791 13970
+ review_scores_communication 1  17929.3 62124 15080
+ review_scores_checkin      1   15746.7 64307 15315
+ review_scores_location     1    7247.0 72806 16144
+ room_type                  3    1593.6 78460 16654
+ price                      1     304.0 79750 16765
+ number_of_reviews          1     475.6 79578 16768
<none>                               80054 16800

Step:  AIC=13538.33
review_scores_rating ~ review_scores_accuracy
```

```
AIC
+ review_scores_cleanliness  1   10528.5 39019 11919
+ review_scores_value        1    8334.6 41213 12265
+ review_scores_communication 1   4548.1 45000 12864
+ review_scores_checkin      1    3681.1 45867 12994
+ review_scores_location     1    1931.7 47616 13276
+ price                      1     309.3 49239 13478
+ room_type                  3     338.4 49210 13518
<none>                               49548 13538
+ number_of_reviews          1       7.8 49540 13546
- review_scores_accuracy     1   30505.6 80054 16791

Step:  AIC=11918.95
review_scores_rating ~ review_scores_accuracy + review_scores_cleanliness
```

```
AIC
+ review_scores_value        1    4271.7 34748 11138
+ review_scores_communication 1   2971.3 36048 11388
+ review_scores_checkin      1    2355.8 36664 11468
+ review_scores_location     1    1372.9 37647 11648
+ room_type                  3     153.7 38866 11866
+ price                      1      95.5 38924 11876
<none>                               39019 11919
+ number_of_reviews          1       7.1 39012 11926
- review_scores_cleanliness  1   10528.5 49548 13538
- review_scores_accuracy     1   12927.8 51947 13843

Step:  AIC=11137.5
review_scores_rating ~ review_scores_accuracy + review_scores_cleanliness +
    review_scores_value
```

```
AIC
+ review_scores_communication 1    1999.5 32748 10742
+ review_scores_checkin      1    1537.1 33211 10794
+ review_scores_location     1     663.8 34084 10971
+ price                      1     218.0 34530 11059
+ room_type                  3     261.4 34486 11112
<none>                               34748 11138
+ number_of_reviews          1      28.9 34719 11141
- review_scores_value        1    4271.7 39019 11892
- review_scores_accuracy     1    6555.0 41303 12280
- review_scores_cleanliness  1    6465.6 41213 12292

Step:  AIC=10742.37
review_scores_rating ~ review_scores_accuracy + review_scores_cleanliness +
    review_scores_value + review_scores_communication
```

```
AIC
+ review_scores_checkin      1     527.5 32221 10588
+ review_scores_location     1     492.9 32255 10648
+ price                      1     213.2 32535 10654
+ room_type                  3     178.1 32570 10661
<none>                               32748 10742
+ number_of_reviews          1      37.4 32711 10743
- review_scores_communication 1    1999.5 34748 11138
- review_scores_value        1    3299.9 36048 11353
- review_scores_accuracy     1    4363.5 37112 11551
- review_scores_cleanliness  1    5838.2 38586 11852

Step:  AIC=10640.52
review_scores_rating ~ review_scores_accuracy + review_scores_cleanliness +
    review_scores_value + review_scores_communication + review_scores_checkin
```

```
AIC
+ review_scores_location       1     462.4 31758 10489
+ price                        1     228.4 31992 10539
+ room_type                    3     167.5 32053 10632
+ number_of_reviews            1      50.6 32170 10639
<none>                                    32221 10640
- review_scores_checkin        1     527.5 32748 10698
- review_scores_communication  1     989.9 33211 10838
- review_scores_value          1    3033.7 35254 11201
- review_scores_accuracy       1    3787.0 36008 11345
- review_scores_cleanliness    1    5661.6 37882 11735

Step:  AIC=10550.83
review_scores_rating ~ review_scores_accuracy + review_scores_cleanliness +
    review_scores_value + review_scores_communication + review_scores_checkin +
    review_scores_location
```

```
AIC
+ price                        1     170.1 31588 10452
+ room_type                    3     129.3 31629 10461
+ number_of_reviews            1      59.4 31699 10547
<none>                                    31758 10551
- review_scores_checkin        1     496.9 32255 10595
- review_scores_location       1     462.4 32221 10640
- review_scores_communication  1     904.5 32663 10733
- review_scores_value          1    2588.3 34347 11023
- review_scores_accuracy       1    3535.2 35294 11208
- review_scores_cleanliness    1    5636.1 37395 11656

Step:  AIC=10523.06
review_scores_rating ~ review_scores_accuracy + review_scores_cleanliness +
    review_scores_value + review_scores_communication + review_scores_checkin +
    review_scores_location + price
```

```
AIC
+ number_of_reviews            1      48.5 31540 10442
- price                        1     170.1 31758 10489
<none>                                    31588 10523
+ room_type                    3      59.0 31529 10537
- review_scores_checkin        1     511.5 32100 10562
- review_scores_location       1     404.0 31992 10601
- review_scores_communication  1     899.5 32488 10706
- review_scores_value          1    2686.2 34275 11009
- review_scores_accuracy       1    3548.6 35137 11178
- review_scores_cleanliness    1    5402.8 36991 11590

Step:  AIC=10521.41
review_scores_rating ~ review_scores_accuracy + review_scores_cleanliness +
    review_scores_value + review_scores_communication + review_scores_checkin +
    review_scores_location + price + number_of_reviews
```

```
AIC
+ room_type                    3      60.0 31480 10429
- price                        1     159.1 31699 10476
<none>                                    31540 10521
- number_of_reviews            1      48.5 31588 10523
- review_scores_checkin        1     523.5 32063 10554
- review_scores_location       1     412.9 31953 10601
- review_scores_communication  1     898.0 32438 10704
- review_scores_value          1    2702.0 34242 11002
- review_scores_accuracy       1    3575.2 35115 11174
- review_scores_cleanliness    1    5435.9 36976 11596

Step:  AIC=10534.91
review_scores_rating ~ review_scores_accuracy + review_scores_cleanliness +
    review_scores_value + review_scores_communication + review_scores_checkin +
    review_scores_location + price + number_of_reviews + room_type


Call:
lm(formula = review_scores_rating ~ review_scores_accuracy +
    review_scores_cleanliness + review_scores_value + review_scores_communication +
    review_scores_checkin + review_scores_location + price +
    number_of_reviews + room_type, data = airbnb_train)

Coefficients:
                (Intercept)       review_scores_accuracy    review_scores_cleanliness
                  24.7536284                   1.8413129                   1.5873319
        review_scores_value  review_scores_communication        review_scores_checkin
                   1.3961638                   1.1090685                   0.8422442
     review_scores_location                       price           number_of_reviews
                   0.4776080                   0.0007713                  -0.0012254
        room_typeHotel room       room_typePrivate room        room_typeShared room
                  -0.1913761                  -0.1834034                  -0.3894232
```

The co-variates given by this fit are as follows:
review_scores_accuracy, review_scores_communication, review_scores_checkin, price, room_type, number_of_reviews, review_scores_cleanliness, review_scores_value, review_scores_location

Hence building a model using the co-variates.

```
Call:
lm(formula = review_scores_rating ~ price + number_of_reviews +
    review_scores_accuracy + review_scores_checkin + review_scores_cleanliness +
    review_scores_communication + review_scores_location + review_scores_value +
    room_type, data = airbnb_train, nbest = 2, nvmax = 9)

Residuals:
     Min       1Q   Median       3Q      Max
-10.1137  -1.2489   0.3009   1.4957  11.0486

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                24.7536284  0.8340064  29.680  < 2e-16 ***
price                       0.0007713  0.0001731   4.455 8.51e-06 ***
number_of_reviews          -0.0012254  0.0003746  -3.271  0.00108 **
review_scores_accuracy      1.8413129  0.0670455  27.464  < 2e-16 ***
review_scores_checkin       0.8422442  0.0802286  10.498  < 2e-16 ***
review_scores_cleanliness   1.5873319  0.0467937  33.922  < 2e-16 ***
review_scores_communication 1.1090685  0.0806037  13.760  < 2e-16 ***
review_scores_location      0.4776080  0.0513873   9.294  < 2e-16 ***
review_scores_value         1.3961638  0.0574683  24.295  < 2e-16 ***
room_typeHotel room        -0.1913761  0.2920388  -0.655  0.51229
room_typePrivate room      -0.1834034  0.0569807  -3.219  0.00129 **
room_typeShared room       -0.3894232  0.1875915  -2.076  0.03794 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.151 on 6804 degrees of freedom
Multiple R-squared:  0.6068,    Adjusted R-squared:  0.6061
F-statistic: 954.4 on 11 and 6804 DF,  p-value: < 2.2e-16
```

Model Observations:

MSE: 2.151 on 6804 degrees of freedom

R-squared: 0.6068

Adjusted R-squared: 0.6061

The linear model built using best subset regression method has the best combination of MSE and Adjusted R-squared. So, we are selecting the Model 3 as it is considered to be more conservative and more interpretable.

```
Call:
lm(formula = review_scores_rating ~ room_type + number_of_reviews +
    price + review_scores_accuracy + review_scores_cleanliness +
    review_scores_checkin + review_scores_communication + review_scores_location +
    review_scores_value, data = airbnb_train, subset = availability_365,
    nbest = 2, nvmax = 9)

Residuals:
    Min      1Q  Median      3Q     Max
-8.0377 -0.7490  0.2984  1.1523 13.0339

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 21.1157940  0.8378415  25.203  < 2e-16 ***
room_typeHotel room          5.5557543  0.3724456  14.917  < 2e-16 ***
room_typePrivate room       -0.1061646  0.0613394  -1.731   0.0835 .
room_typeShared room        -5.0172113  0.1603455 -31.290  < 2e-16 ***
number_of_reviews           -0.0048777  0.0004175 -11.683  < 2e-16 ***
price                        0.0017017  0.0001413  12.040  < 2e-16 ***
review_scores_accuracy       3.1033903  0.0760864  40.788  < 2e-16 ***
review_scores_cleanliness    2.1124615  0.0519566  40.658  < 2e-16 ***
review_scores_checkin        1.4755657  0.0812286  18.166  < 2e-16 ***
review_scores_communication -0.4314776  0.0936710  -4.606 4.19e-06 ***
review_scores_location       0.0459386  0.0506684   0.907   0.3646
review_scores_value          1.3549523  0.0586379  23.107  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.935 on 5680 degrees of freedom
Multiple R-squared:  0.7222,    Adjusted R-squared:  0.7217
F-statistic:  1343 on 11 and 5680 DF,  p-value: < 2.2e-16
```

Model Observations:

MSE: 1.935 on 5680 degrees of freedom

R-squared: 0.7222

Adjusted R-squared: 0.7217

```
[1] 3.74463
[1] 0.7222235
[1] 0.7216855
```

To evaluate how the model performs on future data, we use predict() to get the predicted values from the test set.

```
[1] 6.726473
[1] 1.897327
```

```
[1] 4.62882
```

Comparing the MSE of Test Dataset, which is equal to 4.24, and the MSPE of the Full Data which is 4.23, we can see that the values are almost similar. Hence the variables that we have selected for our model are good estimators of our dependent variable.

Interesting insights from the above analysis helped us understand the New York Airbnb dataset better. Following insights were drawn from it:

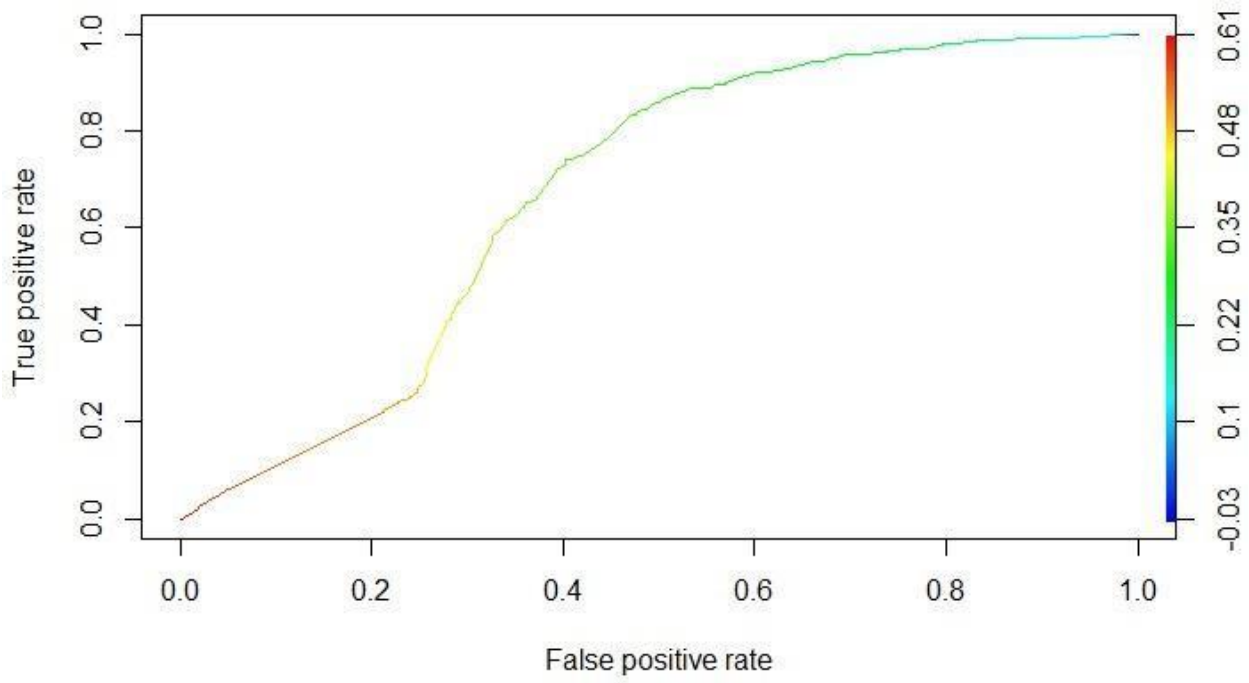
1) Entire home/apt is a common listing in Brooklyn, Manhattan and Staten Island whereas Private room in Bronx and Queens.

2) Shared room is the least common type of listings.
3) Average price of listings is the highest for Manhattan followed by Brooklyn.
4) Bronx has the cheapest listings with an average price of 77.69 USD.
5) Average price is the highest for Hotel room followed by Entire home/apartment.

The factors that impact review scores rating are:
review_scores_accuracy, review_scores_communication, review_scores_checkin, price, room_type, number_of_reviews, review_scores_cleanliness, review_scores_value, review_scores_location

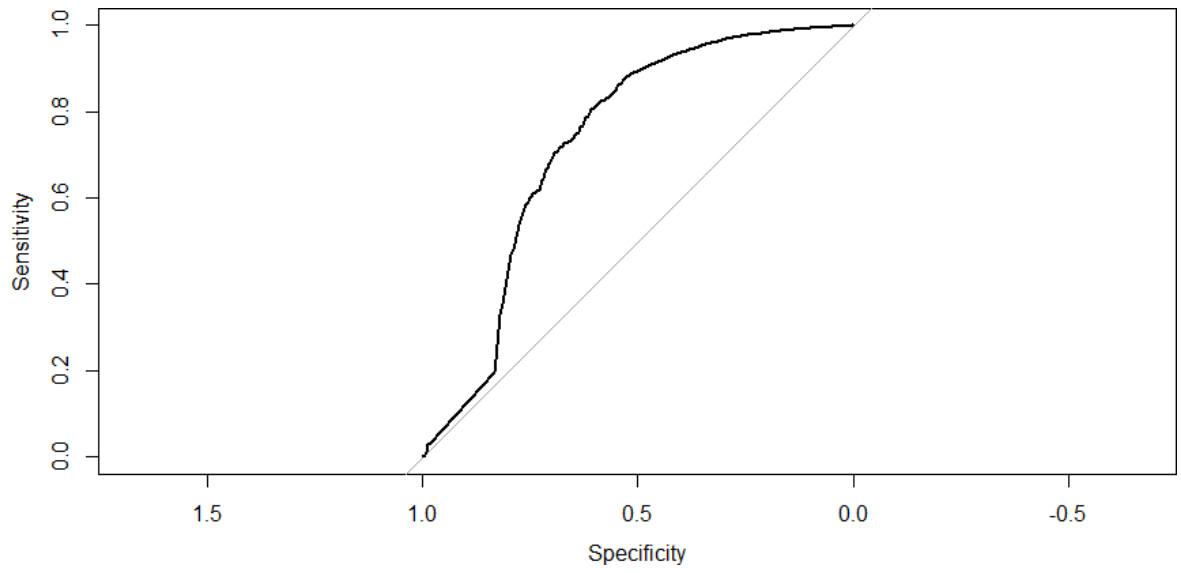
We used different seed numbers for the models to compare them.

LINEAR PROBABILITY MODEL



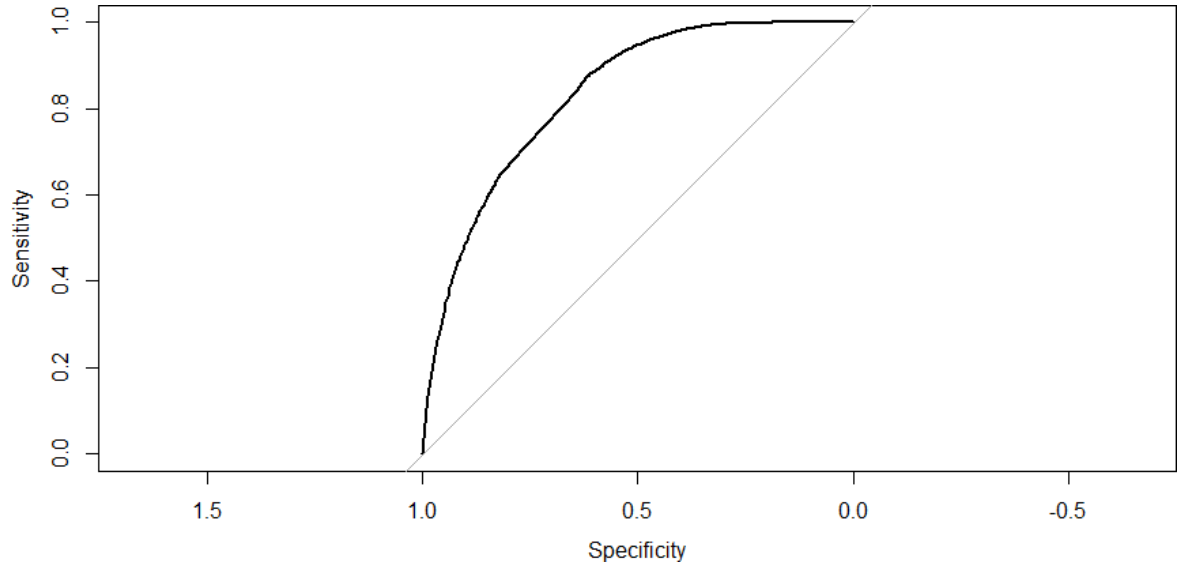
AUC for LPM: 0.67

LOGISTIC MODEL



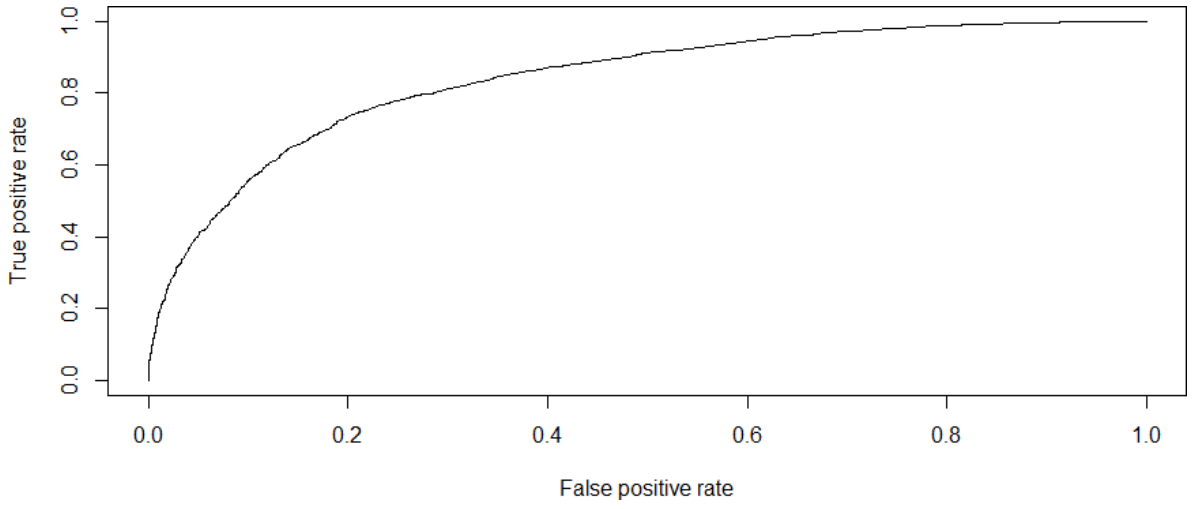
AUC for Logistic Model: 0.7292

MACHINE LEARNING: Random Forest



AUC for random forest: 0.8321

MACHINE LEARNING: Neural Network



AUC for Neural Network: 0.671

Summary of the Project

SUMMARY

We compared the MSE of test dataset and the MSPE of the Full Data and saw that the values are almost similar and therefore variables that we selected for our model are good estimators of our dependent variable. From our project, we concluded that review_scores_accuracy, review_scores_communication, review_scores_checkin, price, room_type, number_of_reviews, review_scores_cleanliness, review_scores_value, review_scores_location.

are the factors that affect the review score rating the most.

To advise our potential hosts we would recommend for them to focus on cleanliness, location, communication, check in ratings, price, room type and collecting reviews for all the visitors. And for a host to eventually and potentially become a superhost, we would recommend them to focus on the following attributes: accommodates, beds, number of reviews, reviews score rating, reviews for cleanliness, check-in, communication, location, value, calculated host listings count.

POTENTIAL SHORTCOMINGS AND FUTURE WORK

- Only the data from 2019 was used, for future work, we can include more cities and years.
- A fast processing computer should be used for machine learning models.
- We will delve deeper into what positively or negatively affects the review scores rating.

Bibliography

- Dataset: insideairbnb.com
- ECO 520 homework codes by Prof Jin Man Lee
- https://www.kaggle.com/xvivancos/tutorial-clustering-wines-with-k-means
- https://daviddalpiaz.github.io/r4sl/logistic-regression.html
- https://www.statmethods.net/management/subset.html
- https://medium.com/@kyawsawhtoon/log-transformation-purpose-and-interpretation-9444b4b049c9
- https://cran.r-project.org/web/packages/pROC/pROC.pdf

Appendix

- R script that was used to generate output