New York City Airbnb Rental - EDA

BACKGROUND INTRODUCTION

Airbnb is an online marketplace for lodging and hospitality where people can list, discover and book unique rentals around the world. San Francisco based Airbnb began operations in 2008 and currently has thousands of employees across the globe supporting property rentals. The company does not own any real estate but simply acts as a broker and makes money in the form of commissions from each of these rentals.

RESEARCH QUESTION

- The data set contains 48,000 airbnb listings in New York City
- It also contains other metrics such as information about hosts, geographical location, type of rental, availability, reviews and price
- The goal of this project is to **predict the price of AirBnB rentals** in New York City based on the available data

DATA Schema:

- Data dictionary for the dataset used in this project is provided in below table

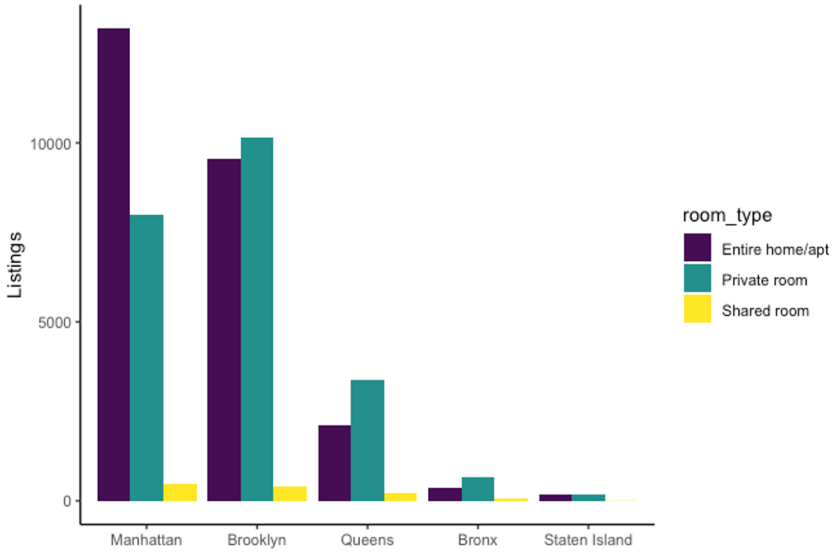
Column Name	Data Type
id	Int
name	Str
host_id	Int
host_name	Str
neighbourhood_group	Str
neighbourhood	Str
latitude	Float
longitude	Float
room_type	Str
price	Float
minimum_nights	Int
number_of_reviews	Int
last_review	Date
reviews_per_month	Float
calculated_host_listings_count	Int
availability_365	Int

DATA: PRE-PROCESSING

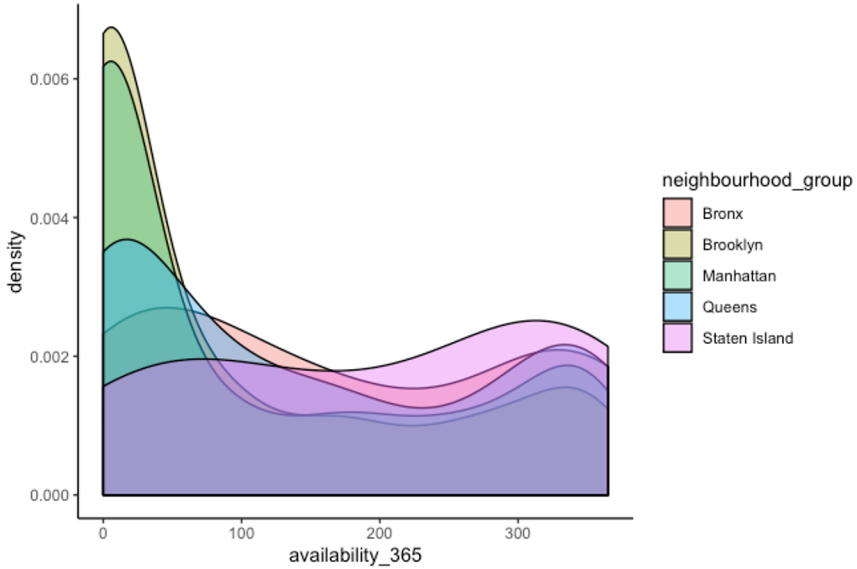
- Variable columns are converted to appropriate data types
- Missing data is handled by backfilling with zeros
- Ensure that all the columns have appropriate values / no outliers

DATA: EXPLORATORY ANALYSIS

Distribution of Listings by Neighbourhood group



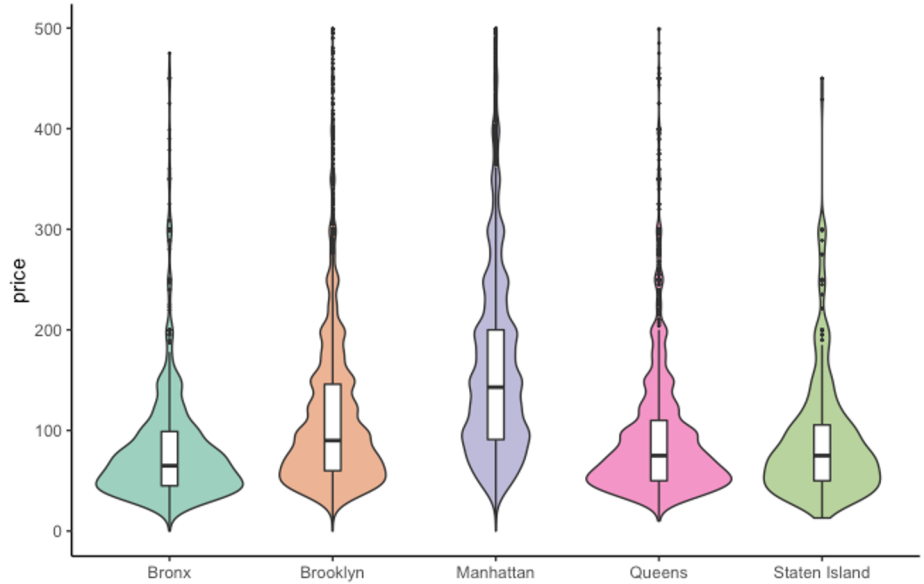
Distribution of Days of Availability by Neighbourhood Group



- Manhattan (44% of listings) and Brooklyn together account for 85% of total listings in NYC
- In terms of availability, Manhattan and Brooklyn have a higher distribution of short-term rentals

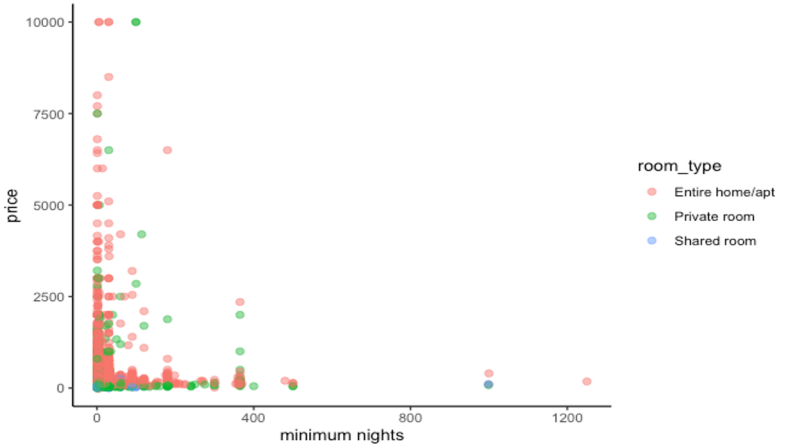
DATA: EXPLORATORY ANALYSIS

Price Variation of Listings by Neighbourhood group

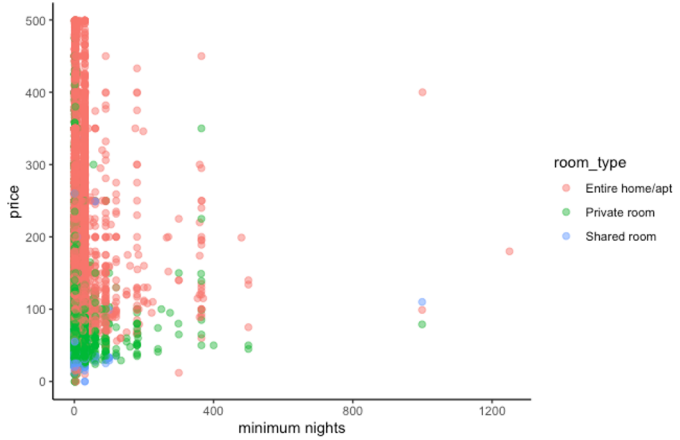


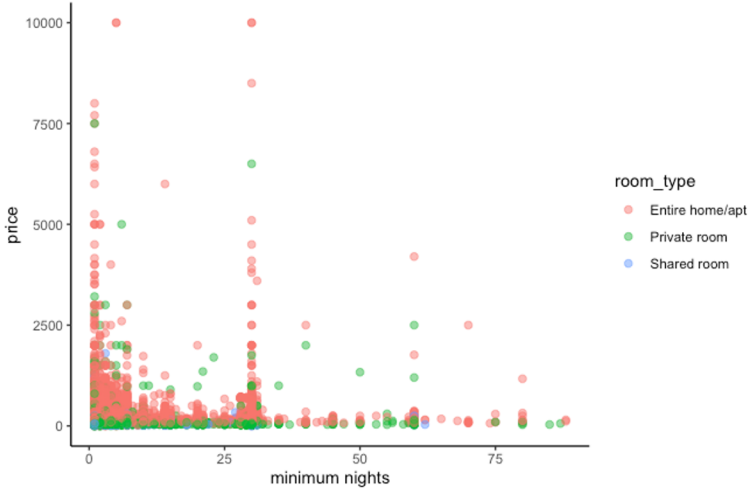
- Manhattan has the highest price range of listings and is 1.5 times more expensive than Brooklyn (median price)
- Although the prices are distributed widely on the lower end for most neighbourhoods, we do see a lot of listings at a range of high price points (long tail at the top)

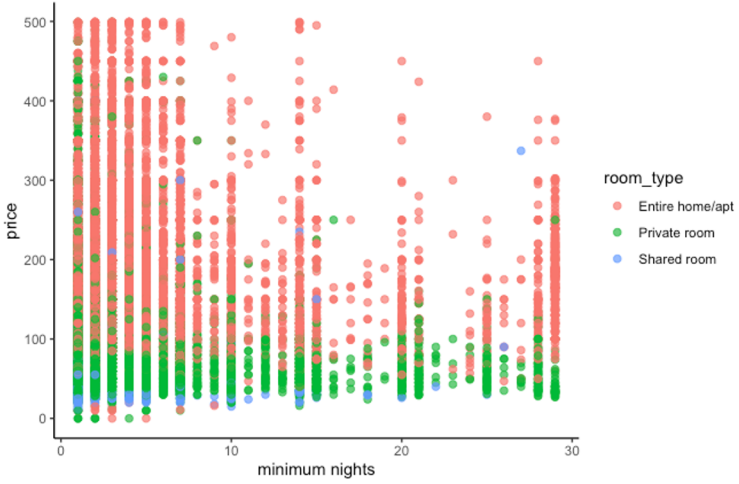
Price Variation of Listings by Minimum nights



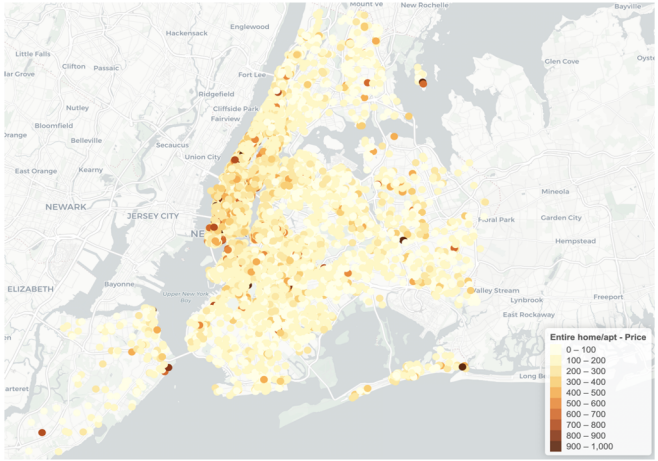
Listings with price < $500



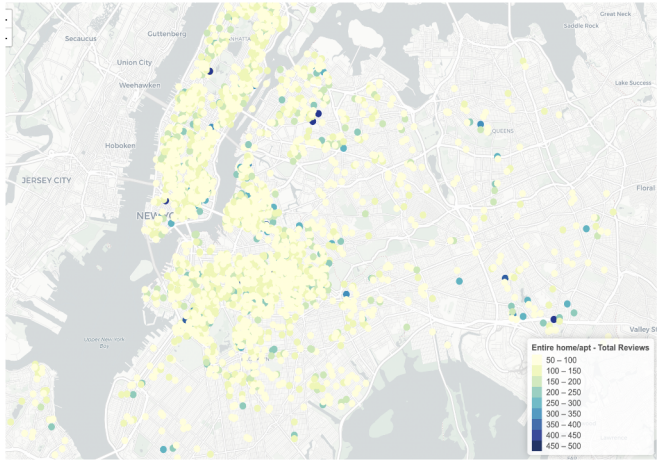




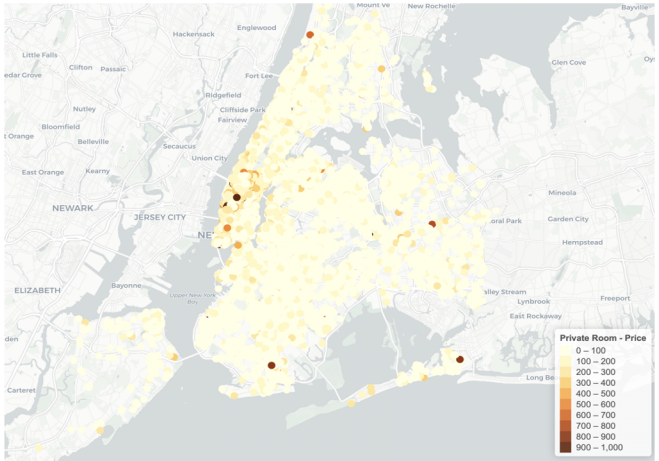
Price Variation by neighbourhood for entire apt



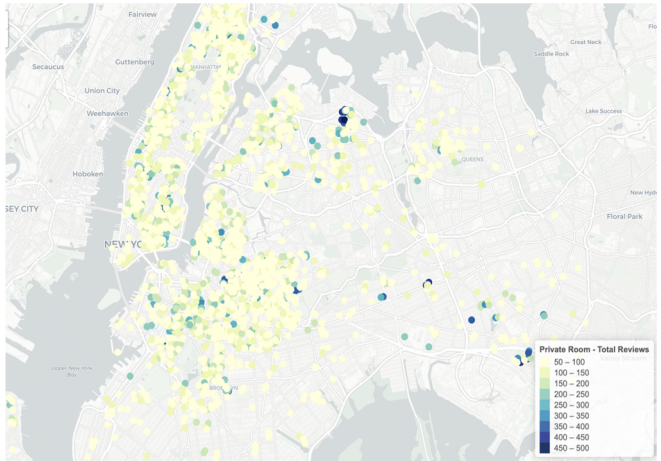
No of reviews by neighbourhood for entire apt



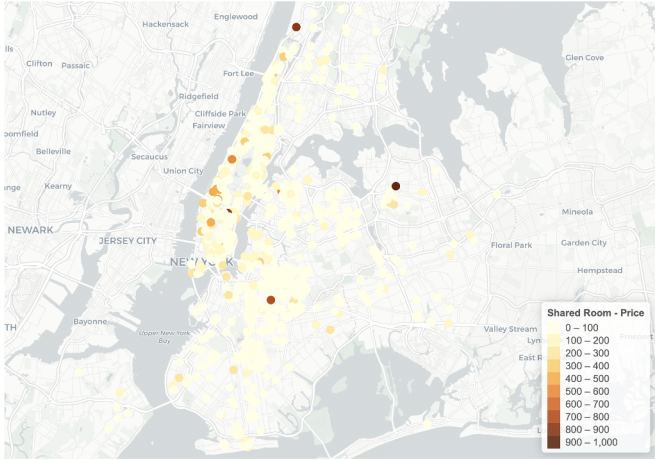
Price Variation by neighbourhood for private room



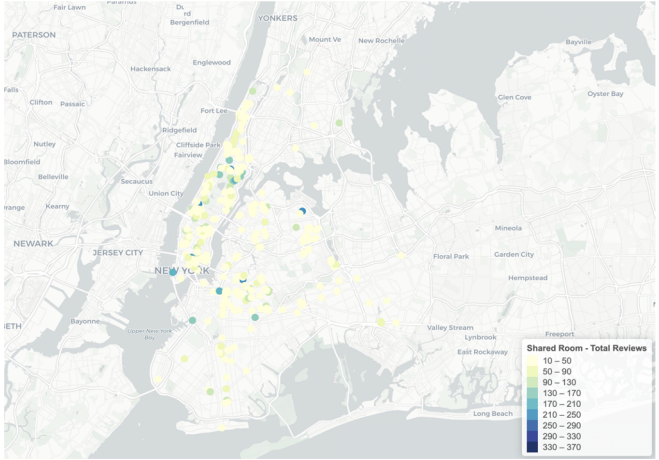
No of reviews by neighbourhood for private room

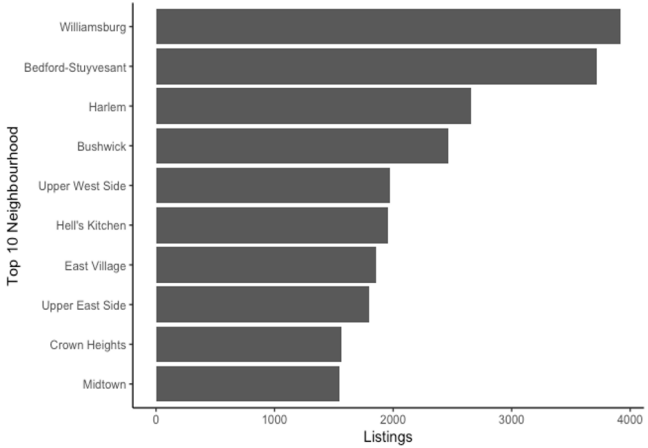


Price Variation by neighbourhood for shared



No of reviews by neighbourhood for shared room





Geographic distribution of listing price supplemented by variation of number of reviews for each listing type helps us identify neighbourhoods with best (*low price, high reviews*) & worst bargains (*high price, low reviews*) within the type.

TECHNICAL APPROACH

- Data cleaning (total observations reduced from 48895 to 46915). Used logarithmic transformation to reduce the skewness in the data.
- Used multiple regression model to prove the conclusion. Applied linear regression model on the specifically chosen columns from the data. Used cook's distance to find the outliers. And upon removed the outliers values from the data.
- Using ANOVA method choose the best model. And conduct experiment to find out how models behaved before and after removal of outliers.
- Finally explained various factors why the variance in price of listing can be explained by 5 predictors mentioned above.

MODEL SUMMARY

Call:
lm(formula = price ~ neighbourhood_group
+room_type+minimum_nights
+number_of_reviews
+calculated_host_listings_count
+availability_365, data = ssAirbnbData)

Residual standard error: 0.507 on 48873 degrees of freedom
Multiple R-squared: 0.4729, Adjusted R-squared: 0.4728
F-statistic: 4385 on 10 and 48873 DF, p-value: < 2.2e-16

Coefficients:

	value Pr(>|t|)
(Intercept)	< 2e-16 ***
neighbourhood_groupBrooklyn	< 2e-16 ***
neighbourhood_groupManhattan	< 2e-16 ***
neighbourhood_groupQueens	1.02e-12 ***
neighbourhood_groupStaten Island	0.968
room_typePrivate room	< 2e-16 ***
room_typeShared room	< 2e-16 ***
minimum_nights	< 2e-16 ***
number_of_reviews	< 2e-16 ***
calculated_host_listings_count	0.505
availability_365	< 2e-16 ***

Summary of model-1 and model-2

Model-1 Call:

```
lm(formula = price ~ neighbourhood_group +
room_type + minimum_nights +
    number_of_reviews + availability_365, data =
ssAirbnbData)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0525 -0.3236 -0.0538  0.2517  5.3652

Coefficients:                          Pr(>|t|)
(Intercept)                          < 2e-16 ***
neighbourhood_groupBrooklyn          < 2e-16 ***
neighbourhood_groupManhattan         < 2e-16 ***
neighbourhood_groupQueens            9.55e-13 ***
neighbourhood_groupStaten Island     0.965
room_typePrivate room                < 2e-16 ***
room_typeShared room                 < 2e-16 ***
minimum_nights                       < 2e-16 ***
number_of_reviews                    < 2e-16 ***
availability_365                     < 2e-16 ***

Residual standard error: 0.507 on 48874 degrees of
freedom
Multiple R-squared:  0.4729,            Adjusted
R-squared:  0.4728
F-statistic:  4872 on 9 and 48874 DF,  p-value: <
2.2e-16
```

Model-2 Call:

```
lm(formula = price ~ neighbourhood_group +
room_type + minimum_nights +
    availability_365, data = ssAirbnbData)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0263 -0.3209 -0.0533  0.2480  5.3613

Coefficients:                          Pr(>|t|)
(Intercept)                          < 2e-16 ***
neighbourhood_groupBrooklyn          < 2e-16 ***
neighbourhood_groupManhattan         < 2e-16 ***
neighbourhood_groupQueens            3.13e-12 ***
neighbourhood_groupStaten Island     0.91
room_typePrivate room                < 2e-16 ***
room_typeShared room                 < 2e-16 ***
minimum_nights                       < 2e-16 ***
availability_365                     < 2e-16 ***

Residual standard error: 0.5082 on 48875 degrees
of freedom
Multiple R-squared:  0.4703,            Adjusted
R-squared:  0.4702
F-statistic:  5423 on 8 and 48875 DF,  p-value: <
2.2e-16
```

ANOVA of model-1 and model-2

```
Analysis of Variance Table

Model 1: price ~ neighbourhood_group + room_type + minimum_nights +
number_of_reviews + availability_365

Model 2: price ~ neighbourhood_group + room_type + minimum_nights +
availability_365

  Res.Df   RSS    Df    Sum of Sq     F      Pr(>F)

1  48874  12562

2  48875  12624   -1    -62.842     244.5   < 2.2e-16 ***
```

Summary of Model-1: before and after outliers removed

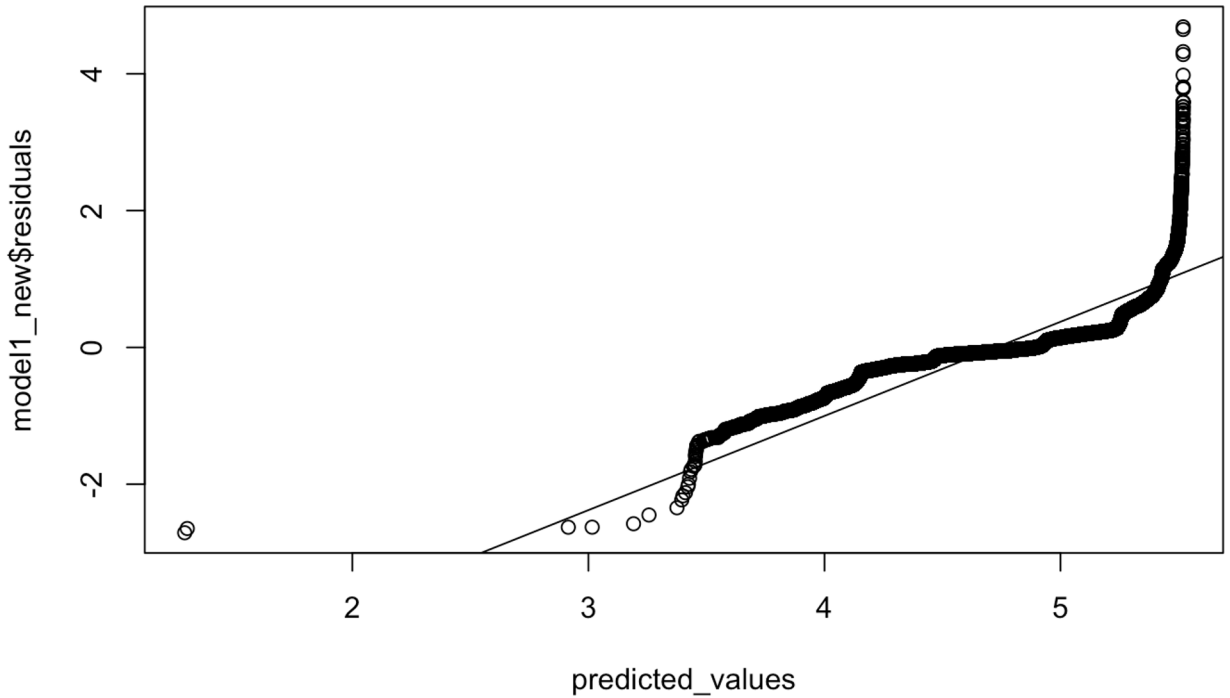
BEFORE Call: lm(formula = price ~ neighbourhood_group + room_type + minimum_nights + number_of_reviews + availability_365, data = ssAirbnbData)

```
Coefficients:                          Pr(>|t|)
(Intercept)                          < 2e-16 ***
neighbourhood_groupBrooklyn          < 2e-16 ***
neighbourhood_groupManhattan         < 2e-16 ***
neighbourhood_groupQueens            9.55e-13 ***
neighbourhood_groupStaten Island     0.965
room_typePrivate room                < 2e-16 ***
room_typeShared room                 < 2e-16 ***
minimum_nights                       < 2e-16 ***
number_of_reviews                    < 2e-16 ***
availability_365                     < 2e-16 ***

Residual standard error: 0.507 on 48874 degrees of
freedom
Multiple R-squared:  0.4729,
Adjusted R-squared:  0.4728
F-statistic:  4872 on 9 and 48874 DF,
p-value: < 2.2e-16 ***
BIC: 72423
AIC: 72326
                        GVIF Df GVIF^(1/(2*Df))
neighbourhood_group 1.057009  4        1.006954
room_type           1.041677  2        1.010260
minimum_nights      1.043278  1        1.021410
number_of_reviews   1.045519  1        1.022506
availability_365    1.081148  1        1.039783
```

AFTER Call: lm(formula = price ~ neighbourhood_group + room_type + minimum_nights + number_of_reviews + availability_365, data = airbnbNoOutliers)

```
Coefficients:                          Pr(>|t|)
(Intercept)                          < 2e-16 ***
neighbourhood_groupBrooklyn          < 2e-16 ***
neighbourhood_groupManhattan         < 2e-16 ***
neighbourhood_groupQueens            2.44e-13 ***
neighbourhood_groupStaten Island     0.371
room_typePrivate room                < 2e-16 ***
room_typeShared room                 < 2e-16 ***
minimum_nights                       < 2e-16 ***
number_of_reviews                    < 2e-16 ***
availability_365                     < 2e-16 ***

Residual standard error: 0.4723 on 46905 degrees of
freedom
Multiple R-squared:  0.507,
Adjusted R-squared:  0.5069
F-statistic:  5359 on 9 and 46905 DF,
p-value: < 2.2e-16
BIC: 62866
AIC: 62770
                        GVIF Df GVIF^(1/(2*Df))
neighbourhood_group 1.058734  4        1.007160
room_type           1.043256  2        1.010643
minimum_nights      1.061314  1        1.030201
number_of_reviews   1.054624  1        1.026949
availability_365    1.094484  1        1.046176
```

Residuals and Fitted-values plot



CONCLUSION

- A Linear regression model was conducted to predict price variations, based on the neighbourhood_group, room_type, minimum_nights, number_of_reviews, availability_365 observations. All the regression assumptions were met. This model is well defined after testing and evaluation.

- A significant regression equation was found with statistical value($F(10,48873)=4385$, p-value < 0.001, R^2 value = 0.4729, BIC = 72432, AIC = 72326). The best fit model was analysed without calculated_host_listings_count and before and after outliers were removed. The statistical result for Model 1 before outliers were removed was ($F(10,48873)=4385$, p-value < 0.001, R^2 value = 0.4729, BIC = 72422, AIC = 72325). Model1 after Outliers removed ($F(10,46905)=5359$, p-value < 0.001, R^2 value = 0.507, BIC = 62866, AIC = 62770).

- When compared the output of Model 1 before and after outliers removed, the results suggested that the 50% of variance in price of listing can be explained by 5 predictors namely neighbourhood_group, room_type, minimum_nights, number_of_reviews, availability_365.