



## Airbnb: New York City

**IDS - 570**

**Statistics for Management**

### **Group 2**

Anurag Sinhal - 660855717

Dikshant Gupta - 675267860

Nikita Maloo - 672689394

Priya Panchaksharappa - 667028888

Lavanya Santhanakrishnan - 662238731

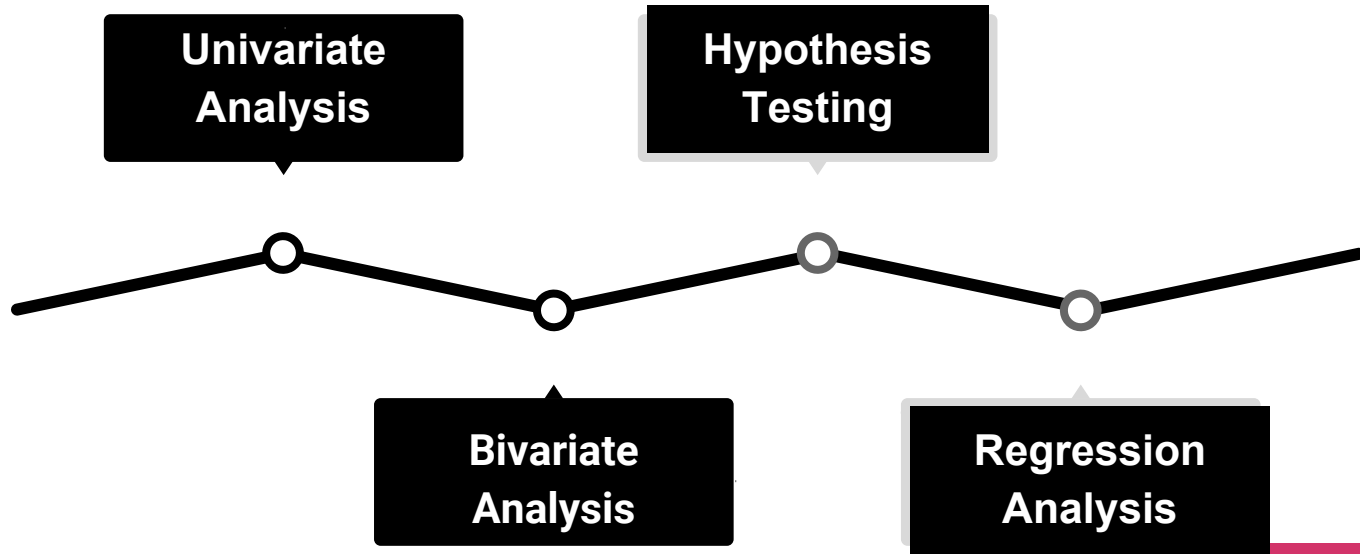
Twinkle Singh - 678538961

# Introduction - Airbnb Company Overview

- Founded in 2008
- Privately owned and operated
- Peer-to-peer online marketplace and homestay network
- People list or rent short-term lodging in residential properties
- Cost of accommodation is set by the property owner
- Receives percentage service fees from both guests and hosts in conjunction with every booking
- Over 2,000,000 listings in 34,000 cities and 191 countries



# Methodology Followed



# Airbnb New York City - Dataset Parameters

The variables that were provided with the dataset

- ID
- Name
- Host\_ID
- Host\_Name
- **Neighborhood\_group**
- **Neighborhood**
- Latitude
- Longitude
- **Room\_type**
- **Price**
- **Minimum\_nights**
- Reviews\_per\_month
- Calculated\_host\_listings\_count
- Availability
- **Number\_of\_reviews**
- **Last\_review**

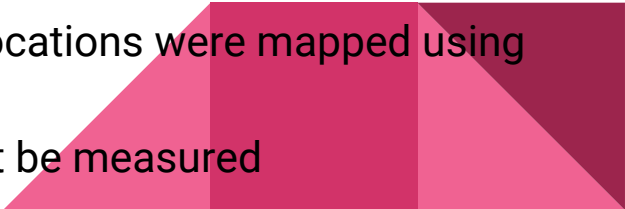
**Note: The highlighted variables are the ones on which our analysis is based primarily upon.**



# Assumptions

- Number of Reviews is assumed as the Number of Bookings for Airbnb
- The month value in Last Review Date variable is assumed as the month in which the booking was done

# Challenges

- New York data has large number of rows, so intense data cleaning was required to draw some meaningful insights
  - Booking data was not explicitly available
  - Tourist spot data was not inherently available (the tourist locations were mapped using internet)
  - The sentiment of the review (positive / negative ) could not be measured
- 

# Research Question

- What are the factors that affect the number of reviews of an airbnb listing in New York City and in what capacity do the factors influence the number of reviews?

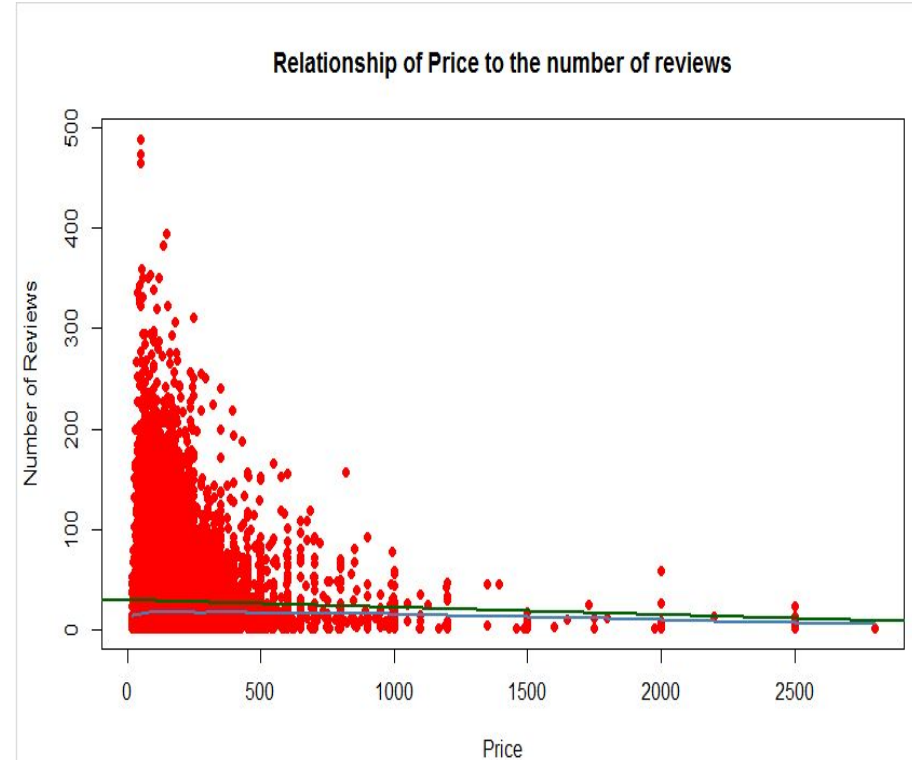


# Bivariate Analysis w.r.t. Dependent Variable

## 1. Number of Reviews Vs. Price

```
Pearson's product-moment correlation  
data: list$price and list$number_of_reviews  
t = -3.7623, df = 21819, p-value = 0.0001688  
alternative hypothesis: true correlation is not equal to 0  
95 percent confidence interval:  
-0.03871715 -0.01219776  
sample estimates:  
cor  
-0.02546193
```

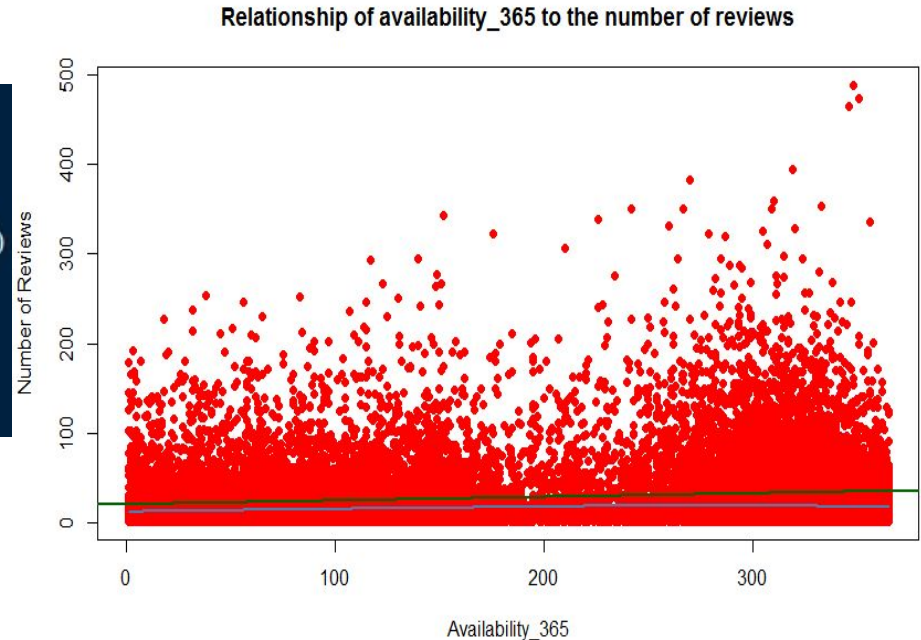
**Result:** 2% Inverse Correlation



# Bivariate Analysis w.r.t. Dependent Variable

## 2. Number of Reviews Vs. Availability 365

```
Pearson's product-moment correlation  
  
data: list$availability_365 and list$number_of_reviews  
t = 19.578, df = 21819, p-value < 0.00000000000000022  
alternative hypothesis: true correlation is not equal to 0  
95 percent confidence interval:  
 0.1183311 0.1444096  
sample estimates:  
      cor  
0.1313931
```



**Result:** 13% Positive Correlation

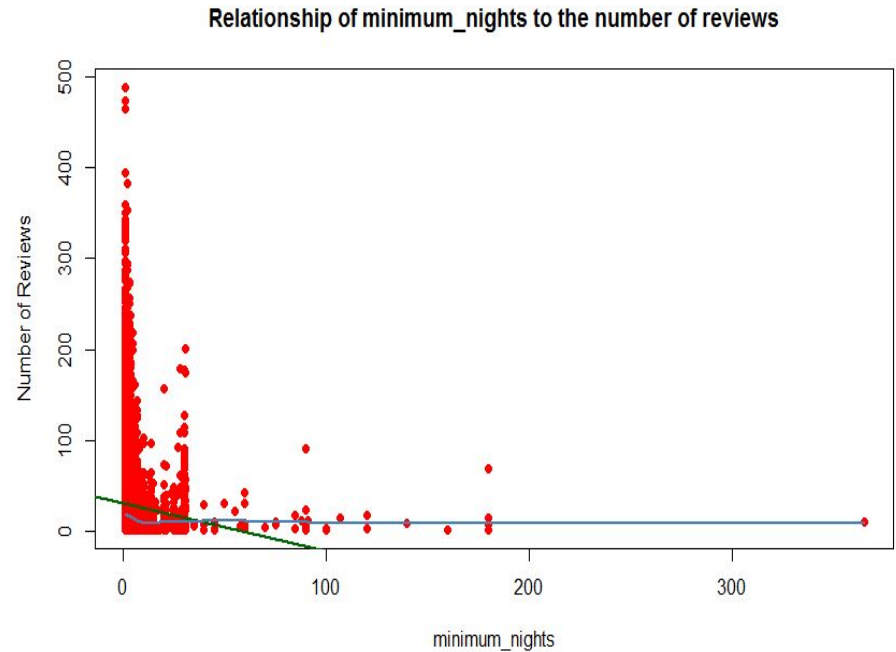


# Bivariate Analysis w.r.t. Dependent Variable

## 3. Number of Reviews Vs. Minimum\_nights

```
Pearson's product-moment correlation  
  
data: list$minimum_nights and list$number_of_reviews  
t = -14.059, df = 21819, p-value < 0.00000000000000022  
alternative hypothesis: true correlation is not equal to 0  
95 percent confidence interval:  
-0.10787969 -0.08158127  
sample estimates:  
cor  
-0.09474701
```

**Result:** 9% Inverse Correlation



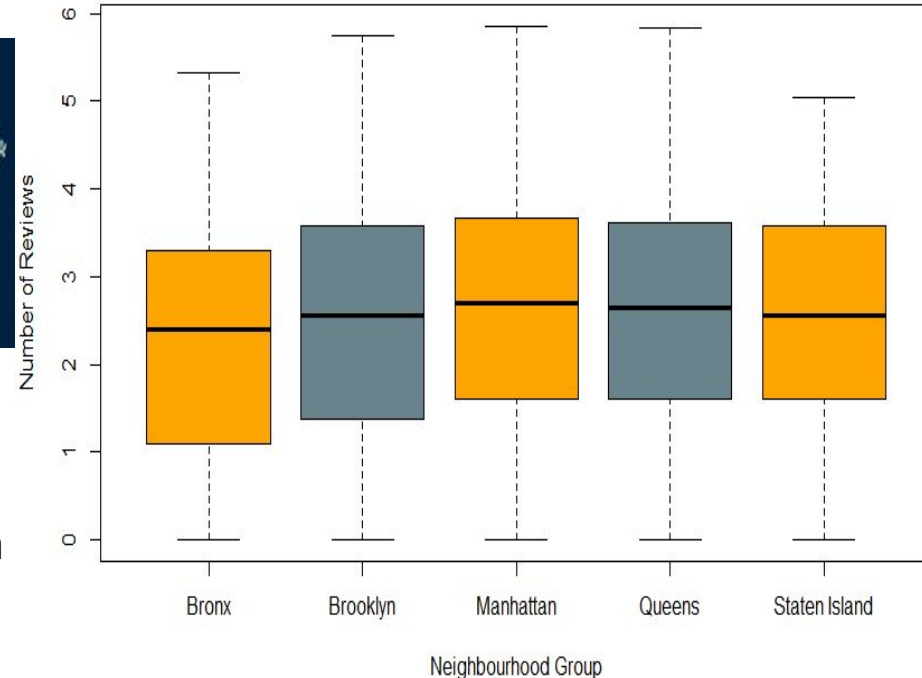
# Bivariate Analysis w.r.t. Dependent Variable

## 4. Number of Reviews Vs. Neighborhood\_Groups

Number of Reviews in Different Neighbourhood groups

```
> summary(tab.aov)
              Df Sum Sq Mean Sq F value    Pr(>F)
neighbourhood_group  4  48479    12120    8.332 0.00000103 ***
Residuals        21816 31731984     1455
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Result:** Number of Reviews are dependent on Neighborhood Groups

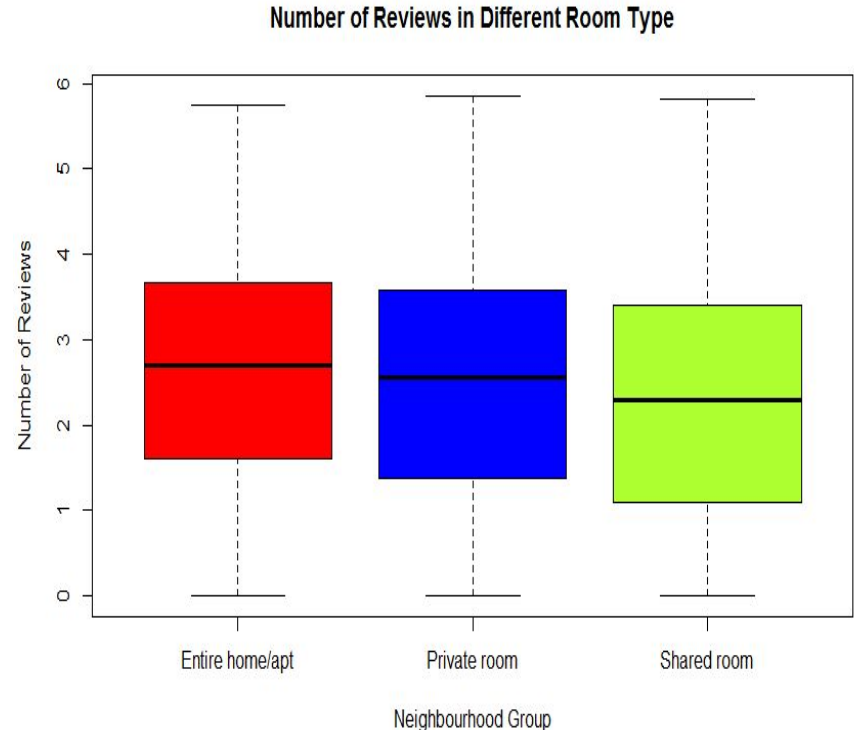


# Bivariate Analysis w.r.t. Dependent Variable

## 5. Number of Reviews Vs. Room Type

```
> summary(list.aov1)
              Df Sum Sq Mean Sq F value    Pr(>F)
room_type      2   40640    20320   13.97 0.000000866 ***
Residuals    21818  31739823     1455
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Result:** Number of Reviews are dependent on Room Types



# Hypotheses

- **The number of reviews of an Airbnb listing in New York City is higher for a listing at lower price**
- **The number of reviews is higher for the listings hosted in Fall or Summer season**
- **The number of reviews for a listing is higher for a neighborhood situated in the vicinity of a tourist destination**



# Hypothesis-1 Analysis

**Hypothesis 1** - The number of reviews of an Airbnb listing in New York City is higher for a listing at lower price

- Initially, price and number of reviews did not display a strong correlation.
- For deeper analysis, price was further sub - categorised

```
data: list$number_of_reviews and list$price
t = -3.9947, df = 21865, p-value = 0.00006498
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.04024567 -0.01375634
sample estimates:
      cor
-0.02700574
```

Price Range	Price Category
Price<=\$100	Economic
\$100< Price <=\$300	Deluxe
Price >\$300	Luxury

# Hypothesis-1 Analysis (Cont.)

## Significant Results of Hypothesis testing on different price categories:

### For Deluxe and Economic price categories:

- a) The correlation coefficient for the price category = **Deluxe** and Neighborhood group = **Staten Island** is -0.2493094
  - b) The correlation coefficient for the price category = **Economic** and Neighborhood group = **Staten Island** is -0.1029916
- Economic and Deluxe price category in Staten Island Neighborhood group, the number of reviews are decreasing with increased prices

### For Luxury price category:

- a) The correlation coefficient for the price category = **Luxury** and Neighborhood group = **Bronx** is around 1
- For Luxury category rooms, the customers are not price centric. Therefore, the number of reviews and price are directly proportional

# Hypothesis-2 Analysis

**Hypothesis 2 - The number of reviews is higher for the listings hosted in Fall or Summer season**

- To analyze the hypothesis, months are divided into seasons as following:

Season	Months
Winter	December, January, February
Spring	March, April, May
Summer	June, July, August
Fall	September, October, November

## Hypothesis-2 (Cont.)

- **ANOVA test between number of reviews and seasons:**

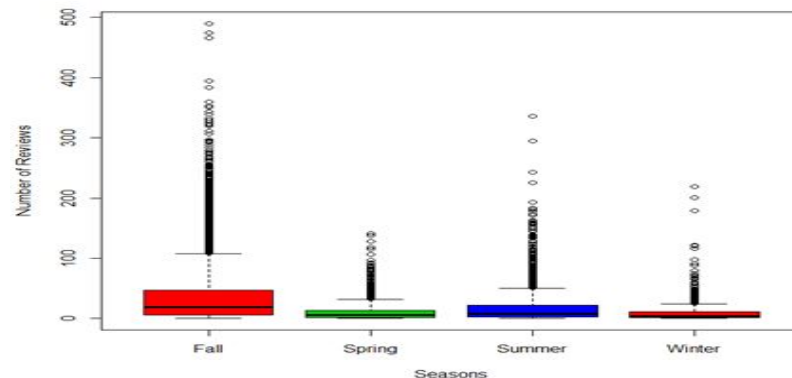
The number of reviews is dependent on the change in seasons

```
              Df Sum Sq Mean Sq F value    Pr(>F)
list$seasons   3  1882680  627560  435.2 <0.0000000000000002 ***
Residuals    21855 31517671    1442

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
8 observations deleted due to missingness
```

- **Peak booking season in New York**

Seasons	Number of Reviews
Fall	35.02987
Spring	11.85124
Summer	17.29261
Winter	10.05087



**Conclusion: Maximum bookings for Airbnb listings are observed for holiday seasons i.e.**

**Fall and Summer**



# Hypothesis-3 Analysis

**Hypothesis 3** - The number of reviews of a listing in a neighborhood is higher if the neighborhood is a tourist destination.

- ANOVA test between number of reviews and neighborhood:  
The number of reviews change with the change in neighborhood

```
              Df Sum Sq Mean Sq F value    Pr(>F)
list$neighbourhood 205  836768    4082  2.715 <0.0000000000000002 ***
Residuals        21661 32569519    1504
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To analyze the number of reviews of a listing near a tourist destination:

- Neighborhood Group – Brooklyn** (Top 3 results):

Neighborhood Name	Tourist Spot	Average Number of Reviews
DUMBO	Brooklyn Bridge	54.94
Coney Island	Coney Island	41.70
South Slope	Brooklyn Art Museum	37.39

## Hypothesis-3 (Cont.)

- Neighborhood Group - Manhattan:

Neighborhood Name	Tourist Spot	Average Number of Reviews
Hell's Kitchen	THE SHED	36.95
Lower East Side	Tenement Museum	35

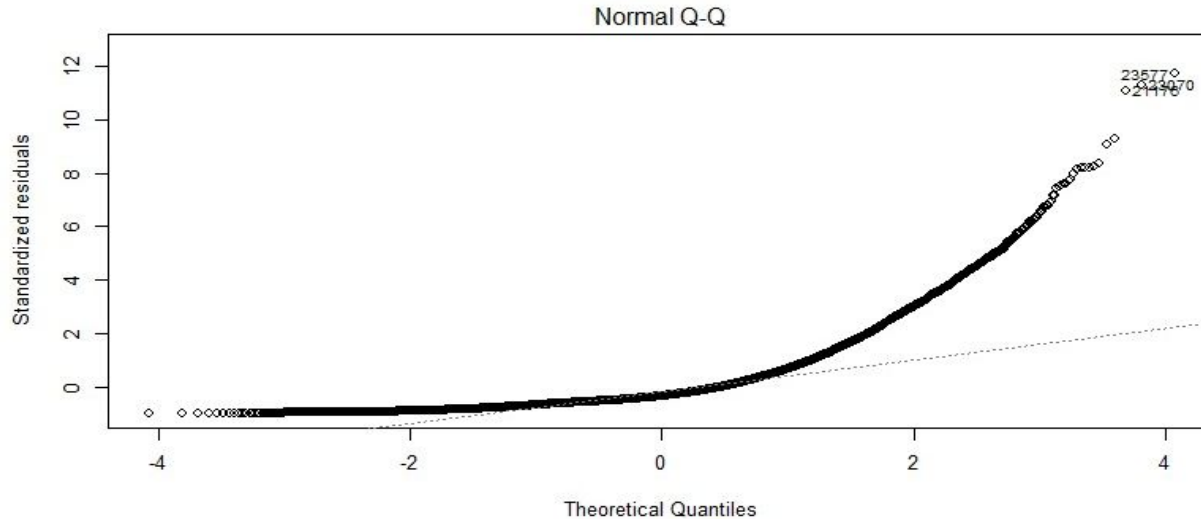
### Conclusion:

The number of reviews for a listing is higher for a neighborhood situated in the vicinity of a tourist destination.

\*tourist spot data searched on the internet.

# Linear Regression Modelling

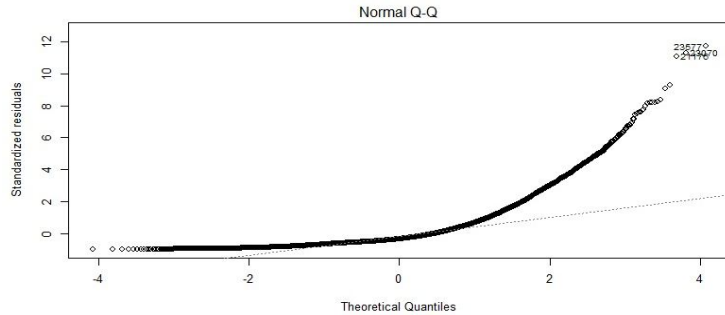
## Model - 1: Number of Reviews Vs Numeric Independent Variables



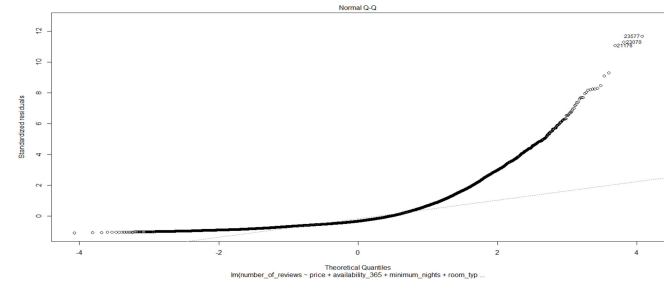
- Adjusted R-squared value is 2.87%
- Factor columns to be added to improve model

# Model - 2

**Number of reviews vs Independent Numeric variables and Factors like Neighborhood group , Room type**



Model 1

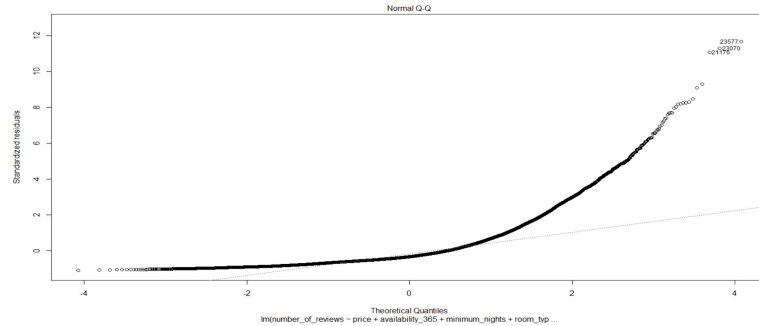


Model 2

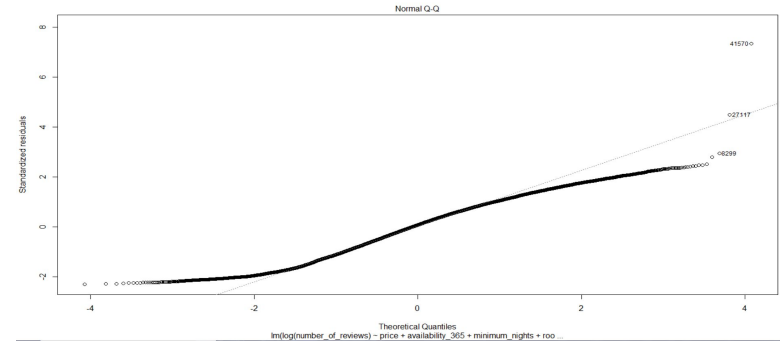
- Enhanced Model 1 by introducing factors in the Linear Model
  - Factors like Neighborhood group, Room Type
- Fitness of Linear Model increased from 2.87 to 3.61%

# Model - 3

## Log transformation of number of reviews



Model 2

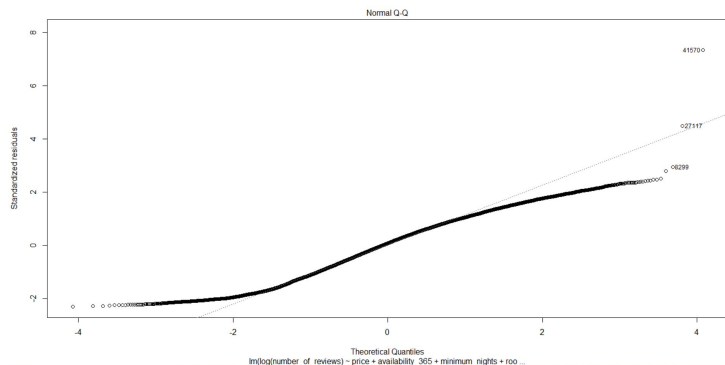


Model 3

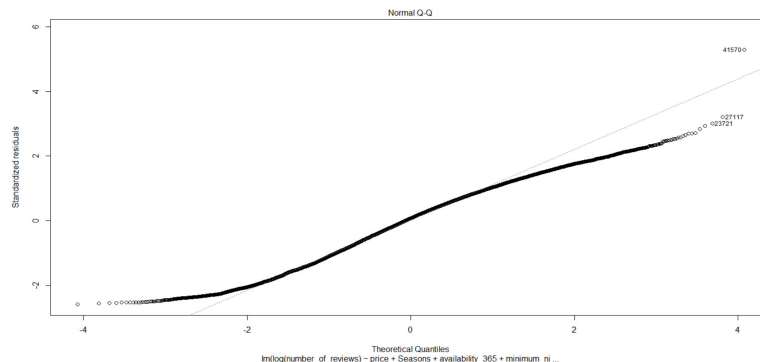
- Log transformation of dependent variable i.e. `number_of_reviews`
- Model fitness improved to 4.45% from 3.6%

# Model - 4

Converting last\_review into factor called Seasons(based on months)



Model 3

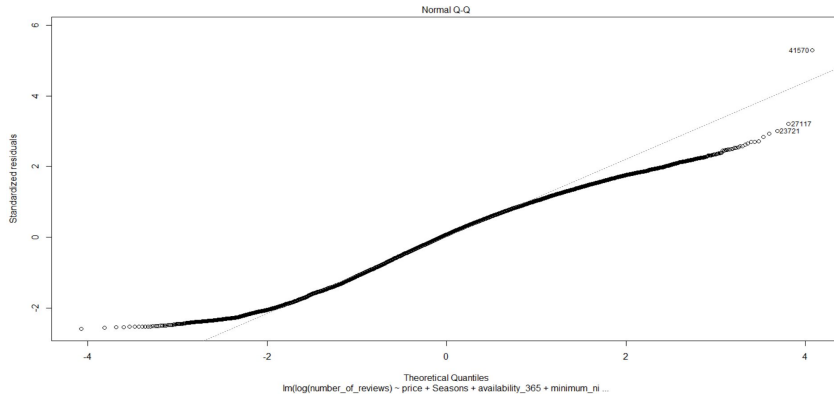


Model 4

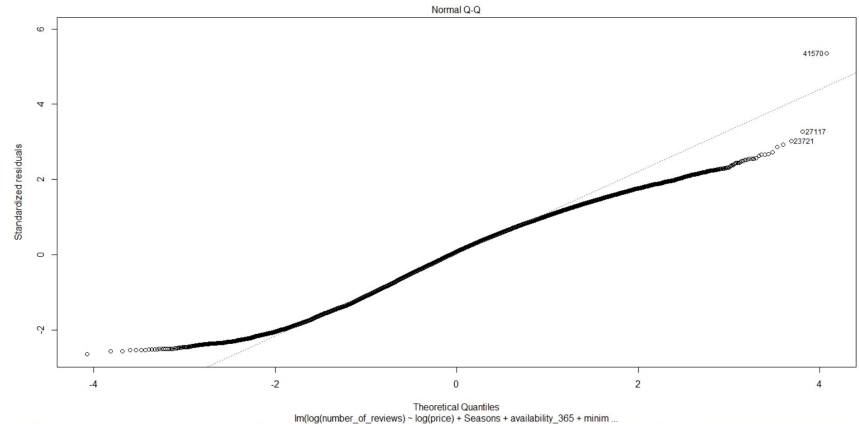
- Seasons factor column introduced with levels - Fall, Winter, Spring, Summer
- Model fitness drastically improved from 4.45% to 12.48 %

# Model - 5

## Log Transformation of Price



Model 4

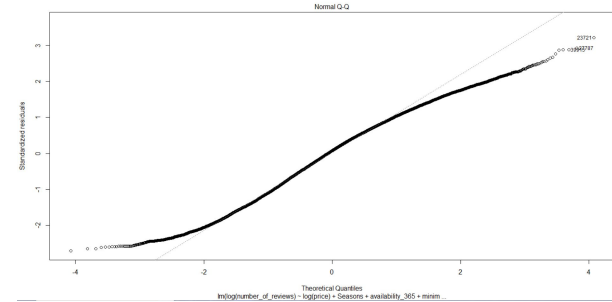
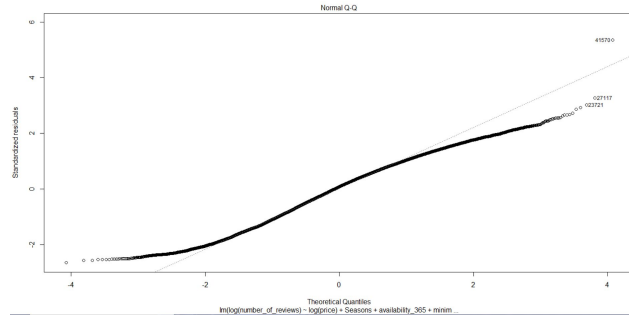


Model 5

- Factored price into 3 categories - Economic, Deluxe, Luxury
  - Fitness increased marginally from 12.48% to 12.50%
- Log of price produced better results

# Model - 6

## Adding Minimum Nights Cat as a factor



- Model fitness increased from 12.5% in Model 5 to 13.8% in Model 6
- Removed outliers for model 7
  - Adjusted R-squared value decreased
  - Model 6 finalised as **BEST LINEAR MODEL**

**\* Model fitness increased by 379.83% from 2.876% for model 1 to 13.88% for model 6**




# Effect of Linear Regression on Hypothesis-1

- Number of reviews will increase with the decrease in the price of a listing
- In the final model, coefficient of the independent variable **Price** (i.e.  $\log(\text{price})$ ) is -0.25149562 which decreased by 0.24 w.r.t. the first model
- Keeping all the other independent variables constant; **if the  $\log(\text{price})$  increases by 1, the  $\log(\text{number of reviews})$  will decrease by 0.25149562**
- **Inference:** Number of reviews and price of a listing are inversely proportional



# Effect of Linear Regression on Hypothesis-2

- Number of reviews will change with the change in the season when the listing was hosted
  - In the final model, coefficient of Season = "Fall" is 0.94785556 and Season= "Summer" is 0.31570821
  - Keeping all the other independent variables constant; **if the hosting in Fall increases by 1, the log(number of reviews) will increase by 0.947**
  - Inference: Number of reviews of a listing is higher in Fall and Summer (Holiday seasons) in comparison to Winter and Spring
- 

# Effect of Linear Regression on Hypothesis-3

- The number of reviews across different neighborhoods is not the same
- In the final model, the coefficient of neighborhood = “Brooklyn” is 0.30664352
- Keeping all the other independent variables constant; **the average difference in log(number of reviews) for a listing in Brooklyn as compared to a listing not in Brooklyn is 0.30664352**
- Inference: Number of reviews for a listing is higher when it is in a neighborhood situated in the vicinity of a tourist destination



# Recommendations

- Airbnb should facilitate customers with discounts and offers during off seasons
- Airbnb should focus more on improving the service quality rather than decreasing the price for a Luxury hosting



**Thank You**

*airbnb*

