# “Airbnb in the Big Apple: A Comprehensive Data Analysis of New York City's Hospitality Landscape in the post Covid-era"

## 1. Introduction

The COVID-19 pandemic has had a significant impact on the hospitality industry, causing a drastic reduction in demand for short-term rental platforms like Airbnb. As travelers became more cautious about their health and safety, they were less likely to stay in shared spaces or with hosts they didn't know. This led to a decrease in occupancy rates and revenue for many Airbnb hosts, making it essential for them to optimize their listings to attract more guests and increase revenue.

By examining the factors that impact Airbnb review ratings in New York, this project aims to provide insights that can help hosts enhance their listings and offer a better experience to guests. This can lead to improved guest satisfaction, higher review ratings, and ultimately increased revenue. In the current climate, it is more important than ever for hosts to focus on delivering a safe and enjoyable experience to guests, and this analysis can provide valuable information to guide their decisions.

As the hospitality industry begins to recover from the impact of COVID-19, the insights gained from this analysis can be particularly helpful in assisting Airbnb hosts to improve their listings and regain guests' trust. By understanding the factors that contribute to guest satisfaction, hosts can take steps to create a more welcoming and safer environment for their guests, which can ultimately lead to higher occupancy rates and revenue.

## 2. Descriptive Analysis

- **Dataset Description**

The dataset we are utilizing is called "Airbnb-NYC-Cleaned" and is obtained from the website Kaggle.com. It consists of 69,305 Airbnb listings, out of which 69,285 are complete. The data pertains to the property listings in five major neighborhoods of New York City, and contains details about their location, price, availability, review ratings, as well as information about the host and the neighborhood. The dataset consists of 22 predictors that are quantitative, binary, or categorical. Our main focus is on the review_rating predictor, while other variables of interest include neighborhood_group, instant_bookable, room_type, price, service_fee, and minimum_nights. The aim of this analysis is to determine the factors that influence review ratings on Airbnb listings and how they can be optimized to improve customer satisfaction and host revenue

```
Rows: 69,305
Columns: 23
$ id                              <int> 1001254, 1002102, 1002403, 1002755, 1003689, 1004098, 1004650, 1005202…
$ name                            <chr> "Clean & quiet apt home by the park", "Skylit Midtown Castle", "THE VI…
$ host_id                         <dbl> 80014485718, 52335172823, 78829239556, 85098326012, 92037596077, 45498…
$ host_identity_verified          <chr> "unconfirmed", "verified", "unconfirmed", "unconfirmed", "verified", "…
$ host_name                       <chr> "Madaline", "Jenna", "Elise", "Garry", "Lyndon", "Michelle", "Alberta"…
$ neighbourhood_group             <chr> "Brooklyn", "Manhattan", "Manhattan", "Brooklyn", "Manhattan", "Manhat…
$ neighbourhood                   <chr> "Kensington", "Midtown", "Harlem", "Clinton Hill", "East Harlem", "Mur…
$ lat                             <dbl> 40.64749, 40.75362, 40.80902, 40.68514, 40.79851, 40.74767, 40.68688, …
$ long                            <dbl> -73.97237, -73.98377, -73.94190, -73.95976, -73.94399, -73.97500, -73.…
$ instant_bookable                <chr> "False", "False", "True", "True", "False", "True", "False", "False", "…
$ cancellation_policy             <chr> "strict", "moderate", "flexible", "moderate", "moderate", "flexible", …
$ room_type                       <chr> "Private room", "Entire home/apt", "Private room", "Entire home/apt", …
$ construction_year               <dbl> 2020, 2007, 2005, 2005, 2009, 2013, 2015, 2009, 2005, 2015, 2004, 2008…
$ price                           <dbl> 966, 142, 620, 368, 204, 577, 71, 1060, 1018, 291, 319, 606, 714, 580,…
$ service_fee                     <dbl> 193.000, 28.000, 124.000, 74.000, 41.000, 115.000, 14.000, 212.000, 20…
$ minimum_nights                  <dbl> 10, 13, 3, 13, 10, 3, 13, 13, 2, 2, 1, 5, 2, 4, 2, 13, 2, 2, 1, 3, 7, …
$ number_of_reviews               <dbl> 9, 45, 0, 270, 9, 74, 49, 49, 430, 118, 160, 53, 188, 167, 113, 27, 14…
$ last_review                     <chr> "2021-10-19", "2022-05-21", "2019-06-14", "2019-07-05", "2018-11-19", …
$ reviews_per_month               <dbl> 0.21, 0.38, 0.79, 4.64, 0.10, 0.59, 0.40, 0.40, 3.47, 0.99, 1.33, 0.43…
$ review_rate_number              <dbl> 4, 4, 5, 4, 3, 3, 5, 5, 3, 5, 3, 4, 4, 4, 3, 3, 3, 5, 3, 5, 3, 3, 4, 3…
$ calculated_host_listings_count  <dbl> 6, 2, 1, 1, 1, 1, 1, 1, 1, 1, 4, 1, 1, 3, 1, 1, 1, 1, 1, 1, 2, 1, 6, 6…
$ availability_365                <dbl> 286, 228, 352, 322, 289, 374, 224, 219, 180, 375, 1, 163, 258, 47, 68,…
$ house_rules                     <chr> "Clean up and treat the home the way you'd like your home to be treate…
```

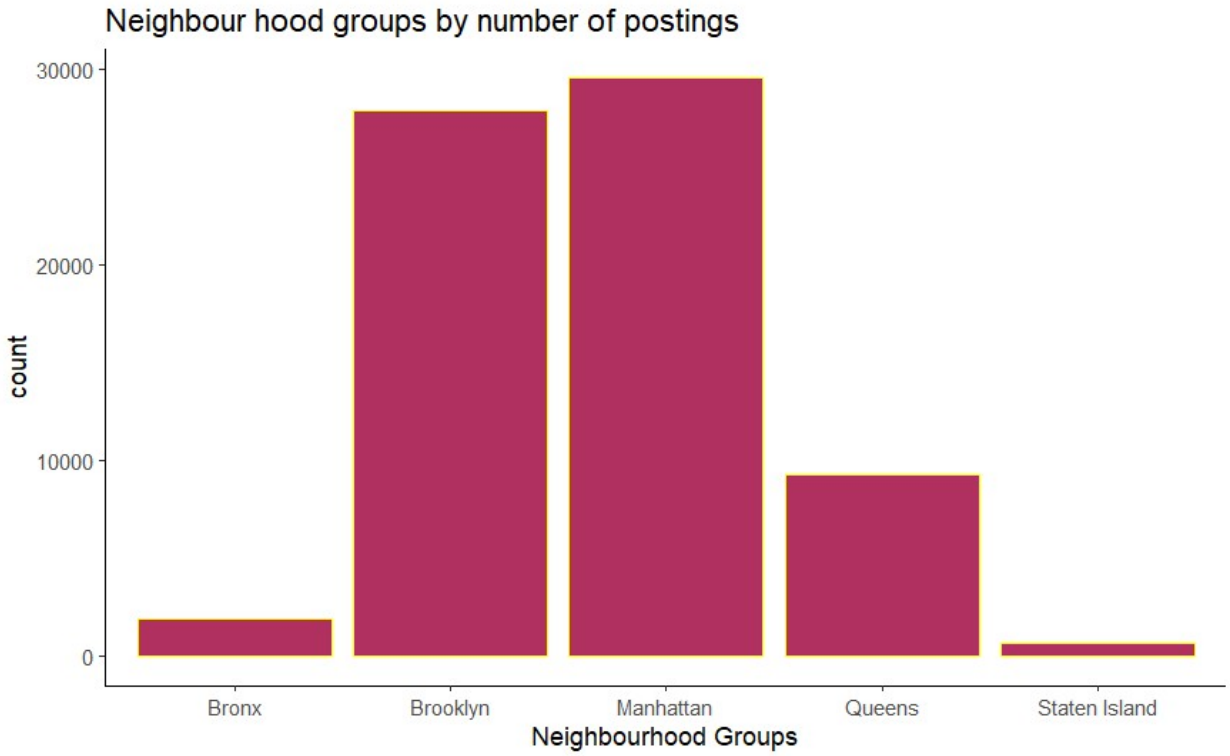
After loading the dataset in and from the head of AB_2021_NYC dataset we can see a number of things. These 23 columns provide a very rich amount of information for deep data exploration we can do on this dataset. We do not see some missing values Meaning this data is clean and read for analysis . Later, we may need to continue with mapping certain values to ones and zeros for predictive analytics. We further explore the variables. We see that neighbourhood_group, neighbourhood, room_type all appear to be categorical variables.

- **Summary Statistics**

```
      id                name              host_id          host_identity_verified  host_name
 Min.   : 1001254   Length:69305       Min.   :1.303e+08   Length:69305           Length:69305
 1st Qu.:10570486   Class :character   1st Qu.:2.460e+10   Class :character       Class :character
 Median :20139636   Mode  :character   Median :4.915e+10   Mode  :character       Mode  :character
 Mean   :20157465                      Mean   :4.929e+10
 3rd Qu.:29708785                      3rd Qu.:7.406e+10
 Max.   :57363551                      Max.   :9.876e+10
 neighbourhood_group neighbourhood          lat             long         instant_bookable
 Length:69305        Length:69305       Min.   :40.50   Min.   :-74.25   Length:69305
 Class :character    Class :character   1st Qu.:40.69   1st Qu.:-73.98   Class :character
 Mode  :character    Mode  :character   Median :40.72   Median :-73.95   Mode  :character
                                        Mean   :40.73   Mean   :-73.95
                                        3rd Qu.:40.76   3rd Qu.:-73.93
                                        Max.   :40.92   Max.   :-73.71
 cancellation_policy  room_type         construction_year     price          service_fee     minimum_nights
 Length:69305        Length:69305       Min.   :2003      Min.   :  50.0   Min.   : 10.0   Min.   : 0.00
 Class :character    Class :character   1st Qu.:2008      1st Qu.: 339.0   1st Qu.: 68.0   1st Qu.: 2.00
 Mode  :character    Mode  :character   Median :2012      Median : 624.7   Median :124.9   Median : 3.00
                                        Mean   :2012      Mean   : 624.7   Mean   :124.9   Mean   : 4.62
                                        3rd Qu.:2017      3rd Qu.: 911.0   3rd Qu.:182.0   3rd Qu.: 6.00
                                        Max.   :2022      Max.   :1200.0   Max.   :240.0   Max.   :13.00
 number_of_reviews last_review        reviews_per_month review_rate_number calculated_host_listings_count
 Min.   :   0      Length:69305       Min.   : 0.010    Min.   :1.000      Min.   :  1.000
 1st Qu.:   1      Class :character   1st Qu.: 0.300    1st Qu.:2.000      1st Qu.:  1.000
 Median :   7      Mode  :character   Median : 0.790    Median :3.000      Median :  1.000
 Mean   :  28                         Mean   : 1.302    Mean   :3.322      Mean   :  8.977
 3rd Qu.:  30                         3rd Qu.: 1.730    3rd Qu.:4.000      3rd Qu.:  3.000
 Max.   :1024                         Max.   :90.000    Max.   :5.000      Max.   :332.000
 availability_365 house_rules
 Min.   :-10.0    Length:69305
 1st Qu.: 18.0    Class :character
 Median :127.0    Mode  :character
 Mean   :153.2
 3rd Qu.:281.0
 Max.   :426.0
```

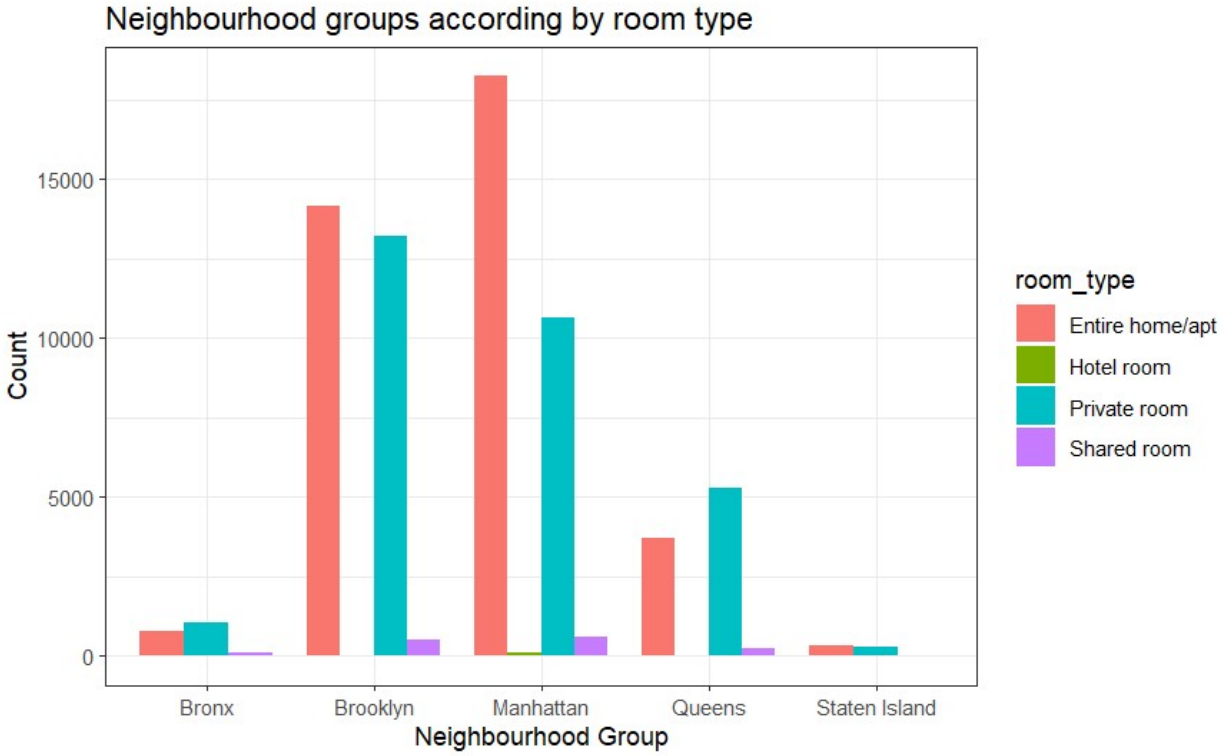
- **Data Visualizations**

Neighborhood groups by number of room postings




Manhattan has the highest number of room postings, followed by Brooklyn with Staten Island having the least postings. This could mean that there are many houses in Manhattan or the condition of the houses are better than the others but this is not something we can infer by looking at the chart.

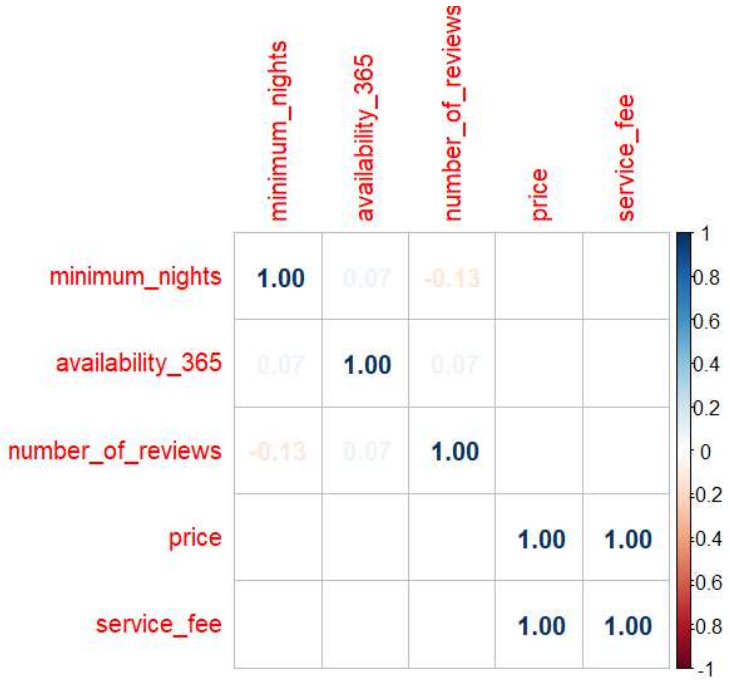
Neighborhood groups according to Room Type




An Entire home/apt is more common in Manhattan than Private room which could infer that people possibly prefer to rent entire home/apt compared to other room types. However, this is not sufficient to answer this question. Shared room is generally low at all the neighborhood groups.

Correlation Analysis

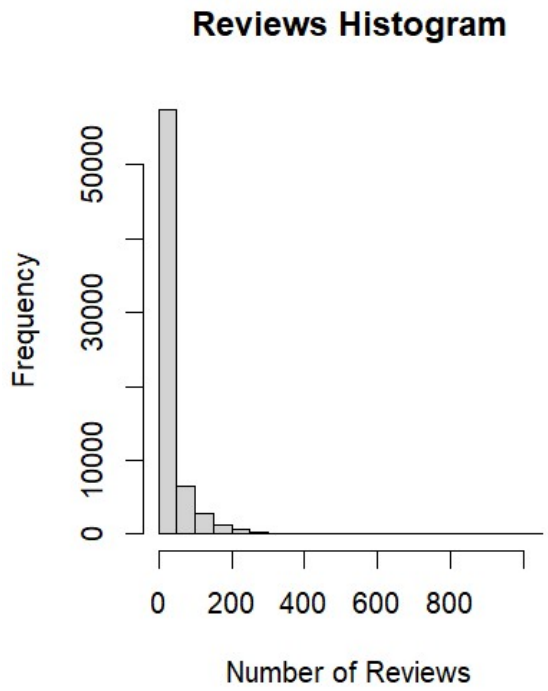
```
> airbnb_data1.cor <- cor(airbnb_data1.mat)
> corrplot(airbnb_data1.cor, order = "hclust", method = "number")
```

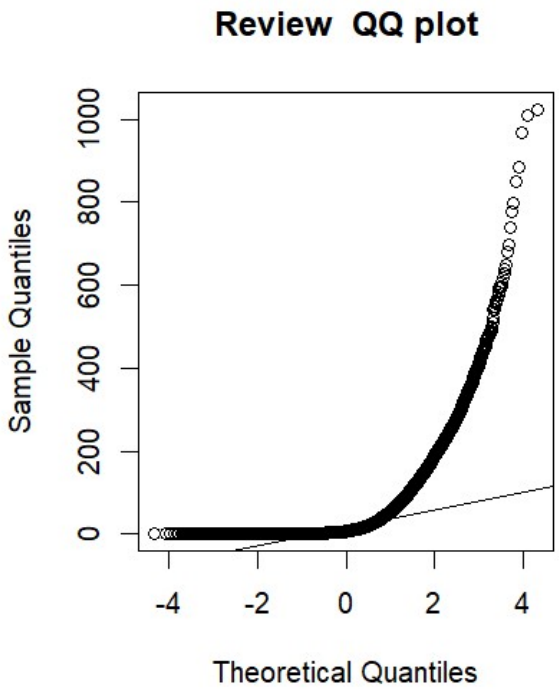



The resulting plot shows a matrix of numbers that represent the strength and direction of the correlations between pairs of variables. The color of each number indicates the sign of the correlation (positive or negative), with blue indicating positive and red indicating negative. The intensity of the color indicates the strength of the correlation, with darker colors indicating stronger correlations.

Normality: For the Response variable number of reviews

```
> par(mfrow = c(1, 2))

> hist(airbnb_data2$number_of_reviews, main = "Reviews Histogram", xlab = "Number of Reviews")

> qqnorm(airbnb_data2$number_of_reviews, main = "Review  QQ plot")

> qqline(airbnb_data2$number_of_reviews)
```

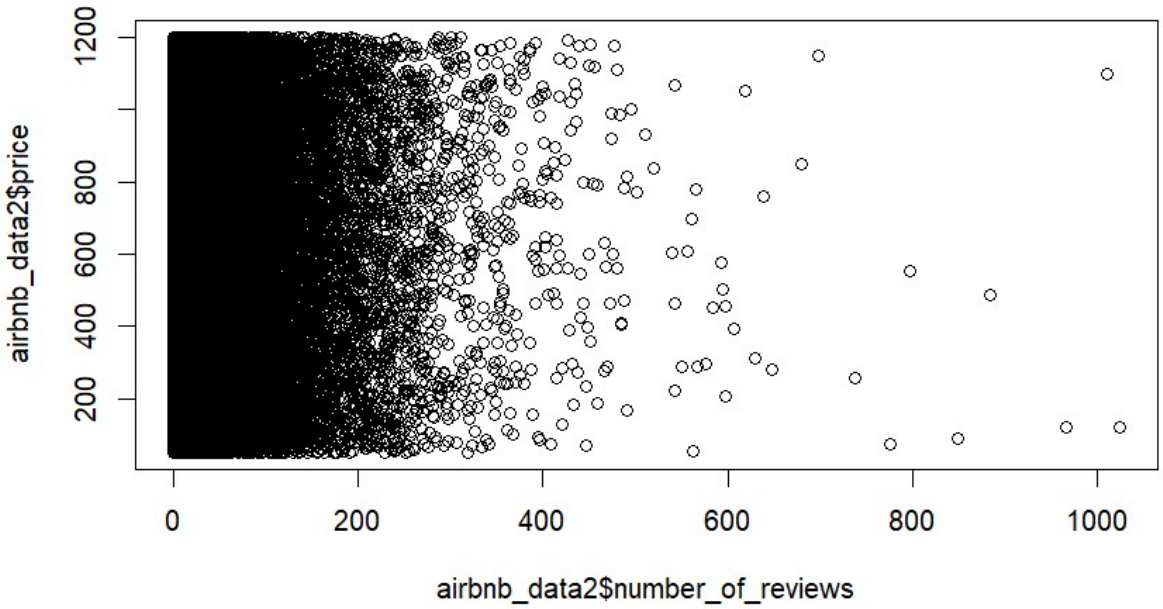



The histogram shows the distribution of the number of reviews for the Airbnb listings. It appears that the majority of listings have a relatively small number of reviews, while a small number of listings have a very large number of reviews. This suggests that there are some highly popular listings on Airbnb that receive a large number of reviews, while the majority of listings receive relatively few reviews.

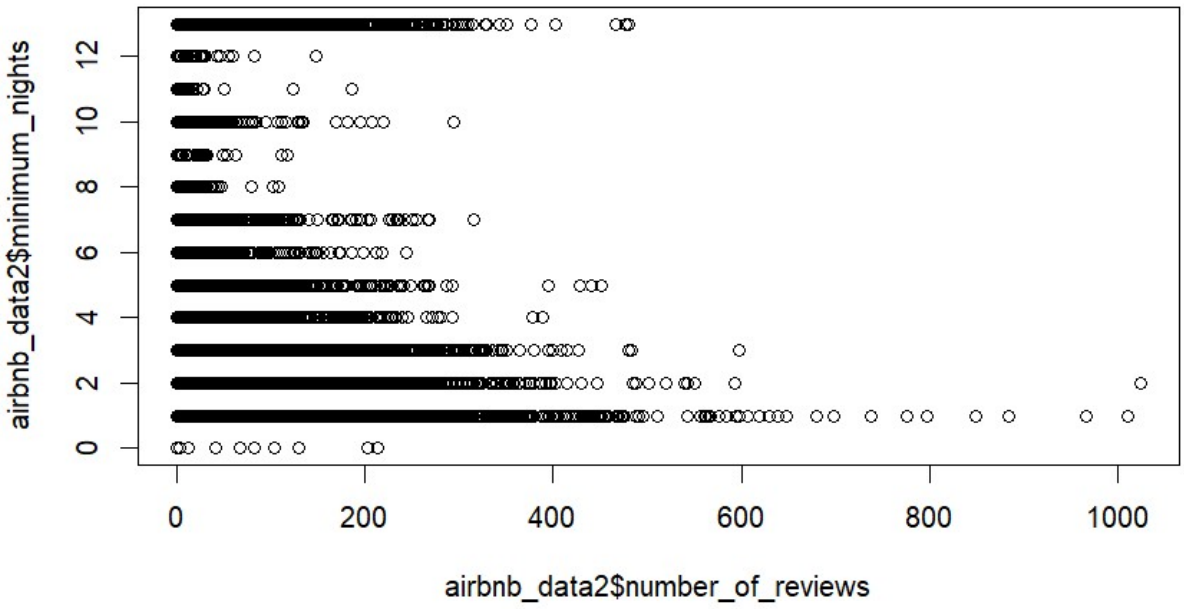
The QQ plot is used to assess whether the distribution of the number of reviews is approximately normal. The plot shows the observed quantiles of the number of reviews against the expected quantiles of a normal distribution. If the data follows a normal distribution, the points on the plot should fall approximately on a straight line. In this case, the plot indicates that the distribution of the number of reviews is not normal, as the points deviate from a straight line in the tails of the distribution. This suggests that the distribution is skewed, with a larger number of listings having a small number of reviews, and a smaller number of listings having a large number of reviews.

Number of reviews vs price Scatter plot




The plot shows the relationship between the number of reviews and price of Airbnb listings. The x-axis represents the number of reviews, and the y-axis represents the price. The scatter plot shows a relatively weak negative correlation between the number of reviews and price, with some outliers. This suggests that, in general, higher priced listings tend to have fewer reviews, while lower priced listings tend to have more reviews. However, there are some exceptions to this trend, as seen by the outliers.

Number of reviews vs Minimum nights




### 3. Hypothesis testing

To analyze the factors that influence the rating criteria offered to Airbnb hosts in New York, we can consider the following variables from the dataset:

- neighbourhood_group
- instant_bookable
- room_type
- price
- service_fee
- minimum_nights

Hypothesis Test:

Null Hypothesis (H0): There is no significant difference in review rates between listings that are instant bookable and those that are not.

Alternative Hypothesis (HA): There is a significant difference in review rates between listings that are instant bookable and those that are not.

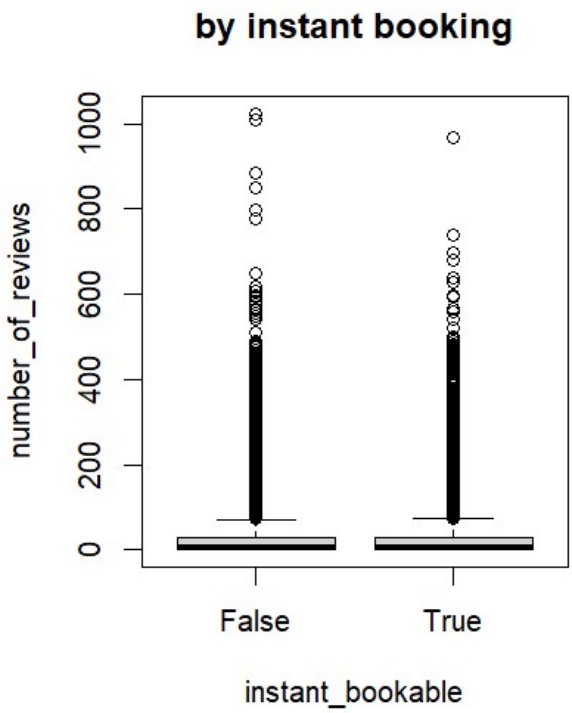
To perform this hypothesis test, we can conduct a two-sample t-test, comparing the mean review rates for instant bookable listings and non-instant bookable listings.

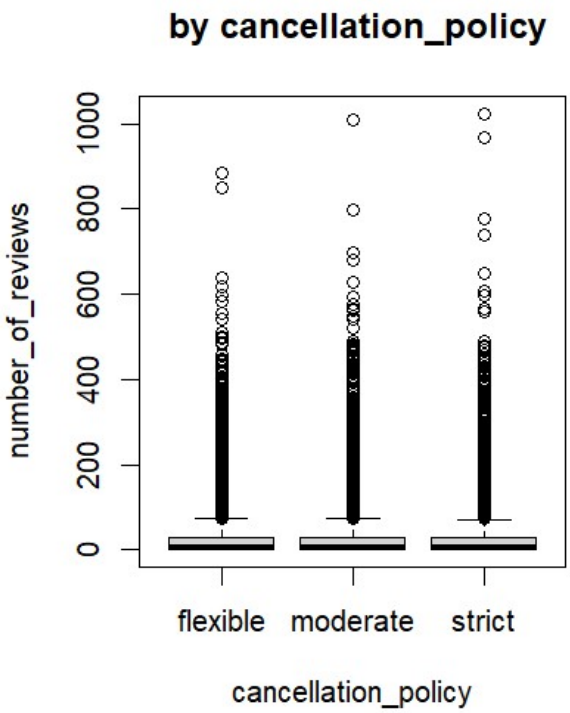
```
        Welch Two Sample t-test

data:  instant$review_rate_number and not_instant$review_rate_number
t = 0.23891, df = 69280, p-value = 0.8112
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01642026  0.02097895
sample estimates:
mean of x mean of y
 3.322783  3.320504

[1] -0.01642026  0.02097895
attr(,"conf.level")
[1] 0.95
```

The test statistic for the Welch Two Sample t-test is 0.23891, with a degree of freedom of 69280 and a p-value of 0.8112. The null hypothesis that there is no difference in mean review rates between instant bookable and non-instant bookable listings cannot be rejected at a significance level of 0.05, as the p-value is greater than the significance level. The 95% confidence interval for the true difference in mean review rates between the two groups is -0.0164 to 0.0210. This means that we are 95% confident that the true difference in mean review rates between instant bookable and non-instant bookable listings falls between these two values. Since the confidence interval includes zero, we cannot conclude that there is a significant difference in mean review rates between the two groups. In other words, there is insufficient evidence to suggest that offering instant booking influences the review rates received by hosts.

## 4. Analysis of Variance (ANOVA)

```
> par(mfrow = c(1, 2))

> boxplot(number_of_reviews~instant_bookable , data = airbnb_data1, main = "by instant booking")

> boxplot(number_of_reviews ~cancellation_policy , data = airbnb_data1, main = " by cancellation_policy")
```




The first boxplot compares the number of reviews between listings with instant booking and those without instant booking. The second boxplot compares the number of reviews between listings with different cancellation policies.

The boxplots show the distribution of the number of reviews within each group, with the median (middle line of the box), the upper and lower quartiles (top and bottom of the box), and whiskers that extend to the highest and lowest values that are within 1.5 times the interquartile range from the box.

Based on these boxplots, it seems that there is not a significant difference in the number of reviews between listings with and without instant booking, as the boxes are very similar in size and the medians are close. However, there appears to be a noticeable difference in the number of reviews between listings with different cancellation policies, as the boxes have different sizes and medians. Specifically, the listings with the flexible cancellation policy seem to have more reviews on average compared to the other policies.

## 5. Regression

*Fitting the Regression model (initial set of predictors)*

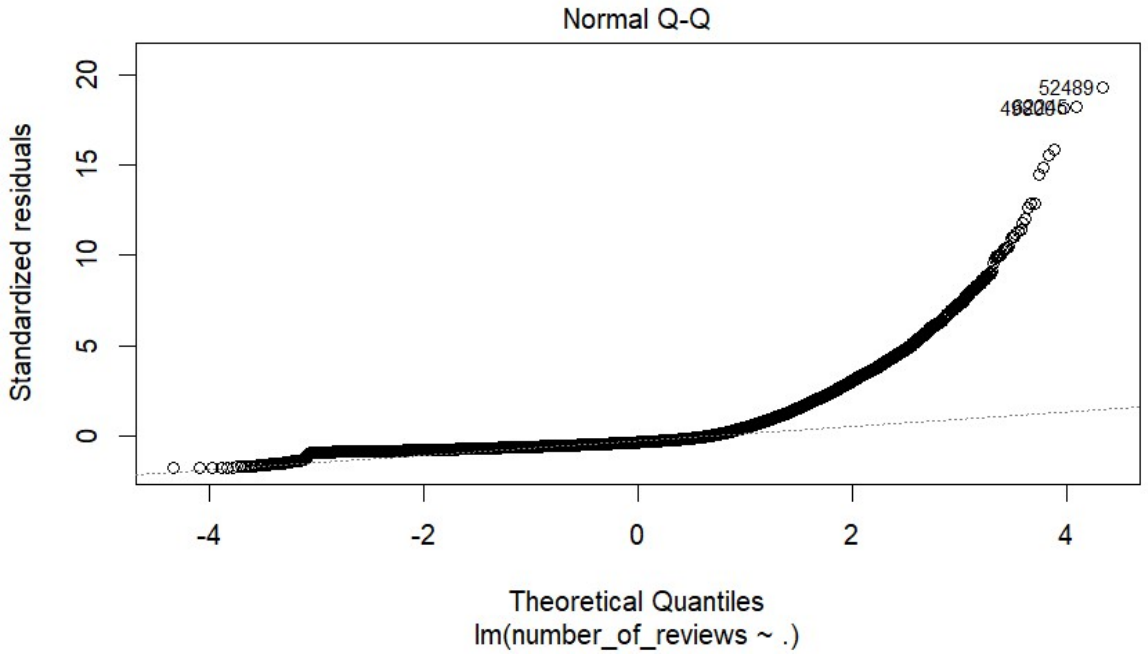
```
Call:
lm(formula = number_of_reviews ~ ., data = airbnb_data1)

Residuals:
   Min     1Q Median     3Q    Max
-89.86 -25.89 -16.69   1.77 987.72

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                   32.444542   1.350662  24.021  < 2e-16 ***
host_identity_verifiedverified -0.161538   0.389642  -0.415 0.678451
neighbourhood_groupBrooklyn     0.576682   1.215969   0.474 0.635318
neighbourhood_groupManhattan   -4.118108   1.216651  -3.385 0.000713 ***
neighbourhood_groupQueens       2.765416   1.288987   2.145 0.031923 *
neighbourhood_groupStaten Island  5.304743   2.330569   2.276 0.022840 *
cancellation_policymoderate     0.106612   0.476357   0.224 0.822909
cancellation_policystrict      -0.608614   0.477952  -1.273 0.202887
room_typeHotel room            51.141339   4.795648  10.664  < 2e-16 ***
room_typePrivate room          -2.729969   0.404982  -6.741 1.59e-11 ***
room_typeShared room          -12.537343   1.367434  -9.169  < 2e-16 ***
price                          -0.000799   0.007694  -0.104 0.917294
service_fee                     0.008666   0.038477   0.225 0.821811
minimum_nights                 -1.579701   0.045616 -34.630  < 2e-16 ***
availability_365                0.032247   0.001460  22.083  < 2e-16 ***
instant_bookableTrue            0.142413   0.389654   0.365 0.714749
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 51.28 on 69289 degrees of freedom
Multiple R-squared:  0.02892,   Adjusted R-squared:  0.02871
F-statistic: 137.6 on 15 and 69289 DF,  p-value: < 2.2e-16
```

The adjusted R-squared value indicates that the model explains only 2.87% of the variance in the number of reviews, which is relatively low.

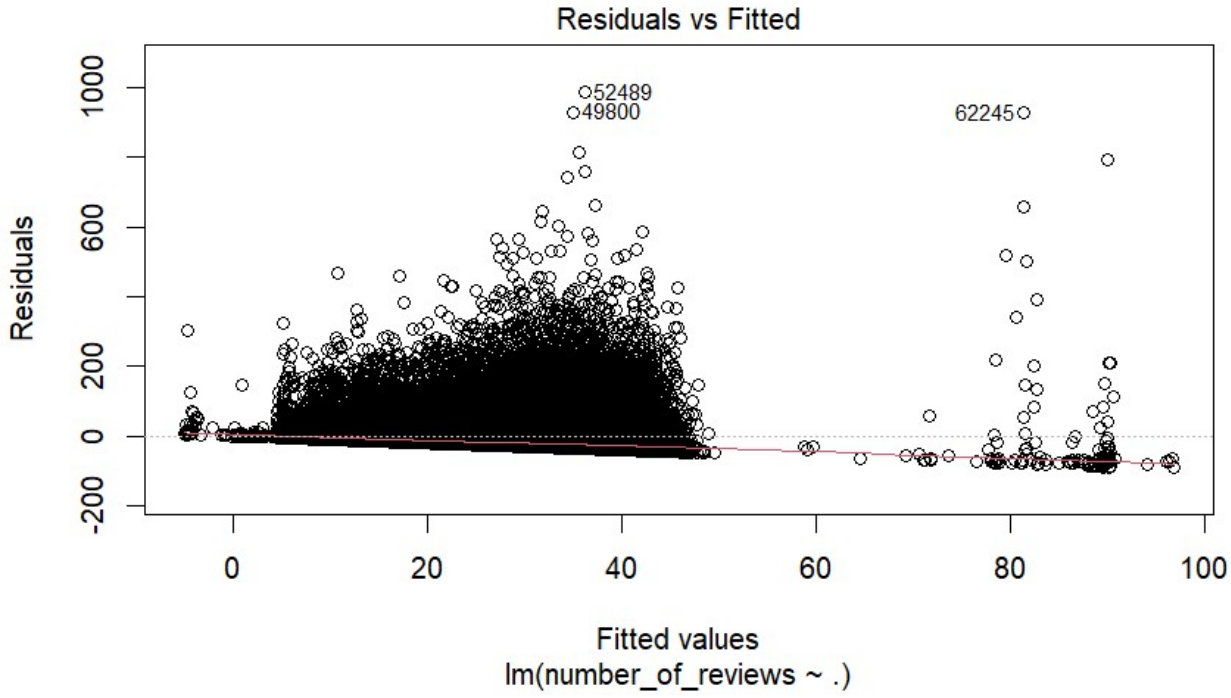
Looking at the individual predictor variables, we can see that some of them are statistically significant at the 5% level ,while others are not. The coefficients for host_identity_verified, cancellation_policymoderate, cancellation_policystrict, instant_bookableTrue, and price are not statistically significant, as their p-values are above the 5% threshold.

On the other hand, the coefficients for neighbourhood_groupManhattan, neighbourhood_groupQueens, neighbourhood_groupStaten Island, room_typeHotel room, room_typePrivate room, room_typeShared room, minimum_nights, and availability_365 are all statistically significant at the 5% level. This means that these variables have a significant linear relationship with the number of reviews.

*Assumtion Test: Residual plot*



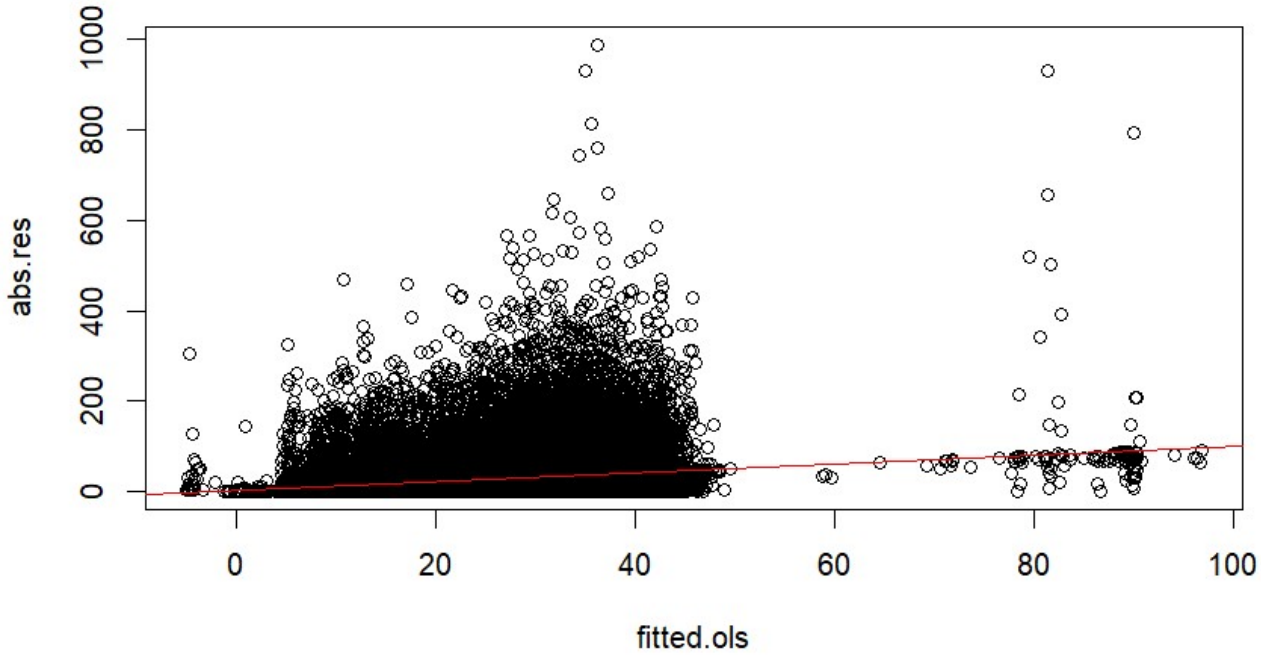

*Inspecting the model for heteroskedasticity*



```
        studentized Breusch-Pagan test

data:  lm.fit.model
BP = 859.84, df = 15, p-value < 2.2e-16
```

The first residual plot clearly shows that the error variance is not even, suggesting that heteroskedasticity may be present. The BP test is also significant, providing evidence of the presence of heteroskedasticity.

*Given that the residuals of the OLS model are heteroskedastic, we fit a Weighted Least Squares WLS model and then fit a linear model to predict abs.res with fitted.ols as the predictor*




This plot is a diagnostic plot for the residuals of a linear regression model. The horizontal axis represents the fitted values (predicted values) of the model, while the vertical axis represents the absolute residuals (the absolute values of the differences between the predicted and observed values).

The red line represents the regression line of the simple linear regression model where the absolute residuals are regressed on the fitted values. The slope of this line indicates the average increase in the absolute residuals for each unit increase in the fitted values.

*Fitting a WLS model*

```
Call:
lm(formula = number_of_reviews ~ ., data = airbnb_data1, weights = wts)

Weighted Residuals:
     Min       1Q   Median       3Q      Max
-1111.44    -3.16     0.04     3.25  1609.23

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -1.145e+02  8.421e+00 -13.595  < 2e-16 ***
host_identity_verifiedverified  1.062e+02  8.426e-01 126.005  < 2e-16 ***
neighbourhood_groupBrooklyn  -6.457e+00  8.245e+00  -0.783 0.433510
neighbourhood_groupManhattan -3.240e+00  8.168e+00  -0.397 0.691644
neighbourhood_groupQueens    -4.930e+00  8.821e+00  -0.559 0.576202
neighbourhood_groupStaten Island  3.099e+00  1.854e+01   0.167 0.867270
cancellation_policymoderate   8.002e+01  9.378e-01  85.327  < 2e-16 ***
cancellation_policystrict     3.714e+01  1.158e+00  32.073  < 2e-16 ***
room_typeHotel room           5.176e+01  8.834e+01   0.586 0.557932
room_typePrivate room         6.335e-01  2.396e+00   0.264 0.791487
room_typeShared room          4.150e+01  2.306e+00  17.998  < 2e-16 ***
price                         8.971e-02  3.888e-02   2.307 0.021032 *
service_fee                  -7.491e-01  1.940e-01  -3.862 0.000113 ***
minimum_nights               -5.439e-01  2.196e-01  -2.477 0.013254 *
availability_365             -5.149e-02  8.844e-03  -5.822 5.83e-09 ***
instant_bookableTrue          1.962e+02  5.862e-01 334.655  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.2 on 69289 degrees of freedom
Multiple R-squared:  0.7027,    Adjusted R-squared:  0.7026
F-statistic: 1.092e+04 on 15 and 69289 DF,  p-value: < 2.2e-16
```

The weighted residuals suggest that the model has some outliers. The coefficients indicate that the intercept and host identity verification have statistically significant positive effects on the number of reviews, while cancellation policy, room type, and price have statistically significant positive effects, and service fee and availability have statistically significant negative effects. The R-squared value of 0.7027 indicates that the model explains 70.27% of the variance in the number of reviews, and the adjusted R-squared value of 0.7026 indicates that the model's complexity does not explain much additional variance. The F-statistic of 1.092e+04 with a very low p-value suggests that the model is a good fit for the data.

## 6. Conclusion

Overall, the analysis of the Airbnb dataset has revealed several interesting insights. Firstly, Manhattan and Brooklyn are the most popular areas for Airbnb listings, with entire homes/apartments being the most common room type in Manhattan. Additionally, while the majority of listings have relatively few reviews, there are some highly popular listings that receive a large number of reviews.

Furthermore, there appears to be a weak negative correlation between the number of reviews and the price of Airbnb listings, with some outliers. Hosts offering instant booking do not seem to receive significantly different review rates compared to those who do not, while listings with more flexible cancellation policies tend to have more reviews

Several predictor variables, such as the borough location, room type, minimum nights, and availability, have a significant linear relationship with the number of reviews. However, some variables, such as host identity verification, cancellation policy, instant booking, and price, do not appear to have a significant impact on the number of reviews.

It is also worth noting that there is evidence of heteroskedasticity in the residual plot of the linear regression model used to predict the number of reviews, which indicates that the variance of the errors is not consistent across different levels of the predictor variables.

Overall, this analysis provides valuable insights for both Airbnb hosts and renters, highlighting the importance of factors such as location, room type, and cancellation policy in determining the popularity of listings.

## 7. References


1. Adding data to the debate 2023, March 6). Retrieved March 25, 2023, from Inside Airbnb: http://insideairbnb.com/new-york-city

2. Zervas, G., & Proserpio, D. (2016, November 18). The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry. Retrieved March 25, 2023, from https://people.bu.edu/zg/publications/airbnb.pdf.