AMS 317 Linear Regression Project

Group 3: Hao Yang Lin, Ariadna Sandoya, Arjun Talapatra, Benjamin Novik

Due Date: November 10, 2024

Introduction

This report investigates the factors influencing Airbnb reviews in New York City. The dataset was taken from Kaggle, which displays 16 variables, and over 40k samples that pertains to NYC Airbnb in 2019. From this, there are scientific questions generated from this dataset, it explores two hypotheses:

1) Room Type Influence on Reviews: Does the type of room (e.g., private room, entire home/apartment) significantly affect the number of reviews per month?

2) Neighborhood Influence on Reviews: Do certain neighborhoods in New York City receive more reviews on average?

To address these hypotheses, the report uses data from the AB_NYC_2019.csv dataset, containing approximately 48,895 observations across multiple variables such as room type, neighborhood, and number of reviews as metrics. This analysis aims to provide insights into how room type and neighborhood characteristics impact review activity, which can inform both hosts and prospective Airbnb guests.

```r
# Loading necessary R library
library(readr)
library(dplyr)
library(ggplot2)
library(car)
library(FSA)
data <- read_csv("airbnb.csv") # Loading Data
```

Data Description

The dataset loaded includes the following columns relevant to our analysis:
- Room Type: Indicates the type of accommodation, with options like “Entire home/apt,” “Private room,” and “Shared room.”
- Neighborhood Group: Represents the borough or neighborhood grouping (e.g., Bronx, Brooklyn, Manhattan, Queens, Staten Island). - Reviews per Month: Measures the average number of reviews a listing receives each month.
- Price: The nightly price of the listing in U.S. dollars (although not a variable for the test, still good to know).
- Other: Various columns such as host_id, latitude, longitude, minimum_nights and availability-related variables that provide context for each listing but are not the main focus in this analysis.

```r
# Summary statistics for variables that are tested for the hypothesis.
summary(data %>% select(room_type, neighbourhood_group, reviews_per_month))
```

```
##   room_type         neighbourhood_group reviews_per_month
##  Length:48895       Length:48895        Min.   : 0.010
##  Class :character   Class :character    1st Qu.: 0.190
##  Mode  :character   Mode  :character    Median : 0.720
##                                         Mean   : 1.373
##                                         3rd Qu.: 2.020
##                                         Max.   :58.500
##                                         NA's   :10052
```

```r
# Check for missing values
missing_values <- sapply(data, function(x) sum(is.na(x)))
print(missing_values)
```

```
##                             id                           name
##                              0                             16
##                        host_id                      host_name
##                              0                             21
##            neighbourhood_group                  neighbourhood
##                              0                              0
##                       latitude                      longitude
##                              0                              0
##                      room_type                          price
##                              0                              0
##                  minimum_nights              number_of_reviews
##                              0                              0
##                     last_review             reviews_per_month
##                          10052                          10052
## calculated_host_listings_count               availability_365
##                              0                              0
```

The resulting output shows us that these following variables have missing values: name, host_name, last_review and reviews_per_month with missing values of 16, 21, 10052 & 10052 respectively. From this, it is shown that the variables, last_review and reviews_per_month are likely correlated. If a listing has no reviews, logically there will not be any reviews for the calculated average. These were replaced with zeroes to accurately reflect the lack of review activity.

```r
# Count rows where number_of_reviews is 0 and reviews_per_month is NA
count_na_reviews_per_month <- sum(is.na(data$reviews_per_month) & data$number_of_reviews == 0)

# Count total rows where number_of_reviews is 0
count_zero_reviews <- sum(data$number_of_reviews == 0)

# Check if all rows with number_of_reviews == 0 have reviews_per_month as NA
all_zero_reviews_are_na <- count_na_reviews_per_month == count_zero_reviews
all_zero_reviews_are_na
```

```
## [1] TRUE
```

```r
# Replace NA values in reviews_per_month with 0 where number_of_reviews is 0
data <- data %>%
  mutate(reviews_per_month = ifelse(is.na(reviews_per_month) &
                                      number_of_reviews == 0, 0, reviews_per_month))
# Check the updated missing values
missing_values <- sapply(data, function(x) sum(is.na(x)))
print(missing_values)
```

```
##                             id                           name
##                              0                             16
##                        host_id                      host_name
##                              0                             21
##           neighbourhood_group                  neighbourhood
##                              0                              0
##                       latitude                      longitude
##                              0                              0
##                      room_type                          price
##                              0                              0
##                minimum_nights              number_of_reviews
##                              0                              0
##                   last_review             reviews_per_month
##                          10052                              0
## calculated_host_listings_count               availability_365
##                              0                              0
```

As observed after the data cleaning process, the majority of variables do not have any missing values anymore aside from last_review, name and host_name. Because these three variables does not pertain any importance to any interest in our analysis, the values can be left as missing.

Data Exploration

```r
# Summary statistics for room types
data %>%
  group_by(room_type) %>%
  summarise(
    count = n(),
    avg_price = mean(price, na.rm = TRUE),
    avg_reviews_per_month = mean(reviews_per_month, na.rm = TRUE)
  )
```

```
## # A tibble: 3 x 4
##   room_type       count avg_price avg_reviews_per_month
##   <chr>           <int>     <dbl>                 <dbl>
## 1 Entire home/apt 25409     212.                   1.05
## 2 Private room    22326      89.8                  1.14
## 3 Shared room      1160      70.1                  1.07
```

```r
# Check for the mean, median, and sd of reviews per month
summary_stats <- data %>%
  group_by(room_type) %>%
```

```
  summarise(
    mean_reviews = mean(reviews_per_month),
    median_reviews = median(reviews_per_month),
    sd_reviews = sd(reviews_per_month)
  )
summary_stats
```

```
## # A tibble: 3 x 4
##   room_type       mean_reviews median_reviews sd_reviews
##   <chr>                  <dbl>          <dbl>      <dbl>
## 1 Entire home/apt         1.05          0.35        1.49
## 2 Private room            1.14          0.4         1.72
## 3 Shared room             1.07          0.405       1.52
```

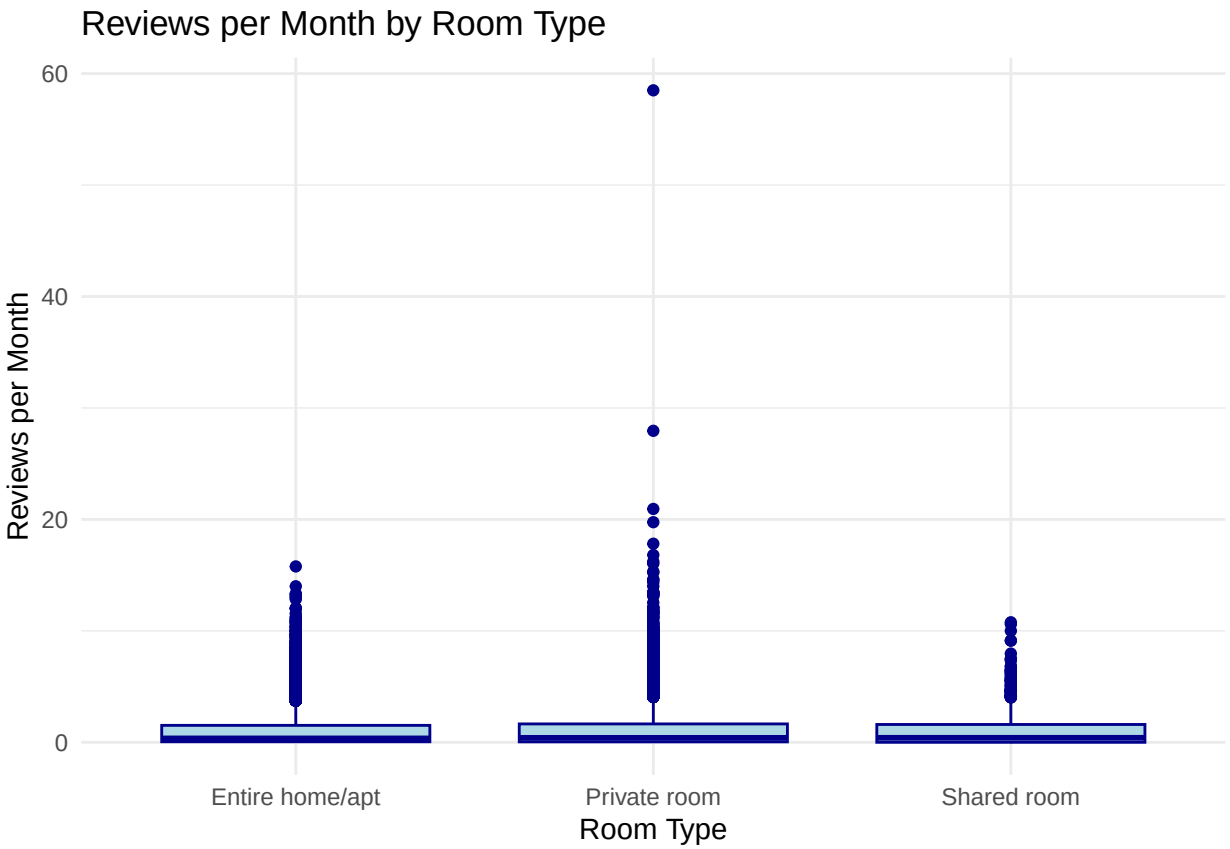
Summary for Distribution of Room Types and Reviews per Month.

The dataset includes three types of rooms: "Entire home/apt," "Private room," and "Shared room." The data distribution is as follows:

- Entire home/apt: 25,409 listings with an average monthly review count of 1.05. Median reviews of 0.350 and standard deviation of 1.486.

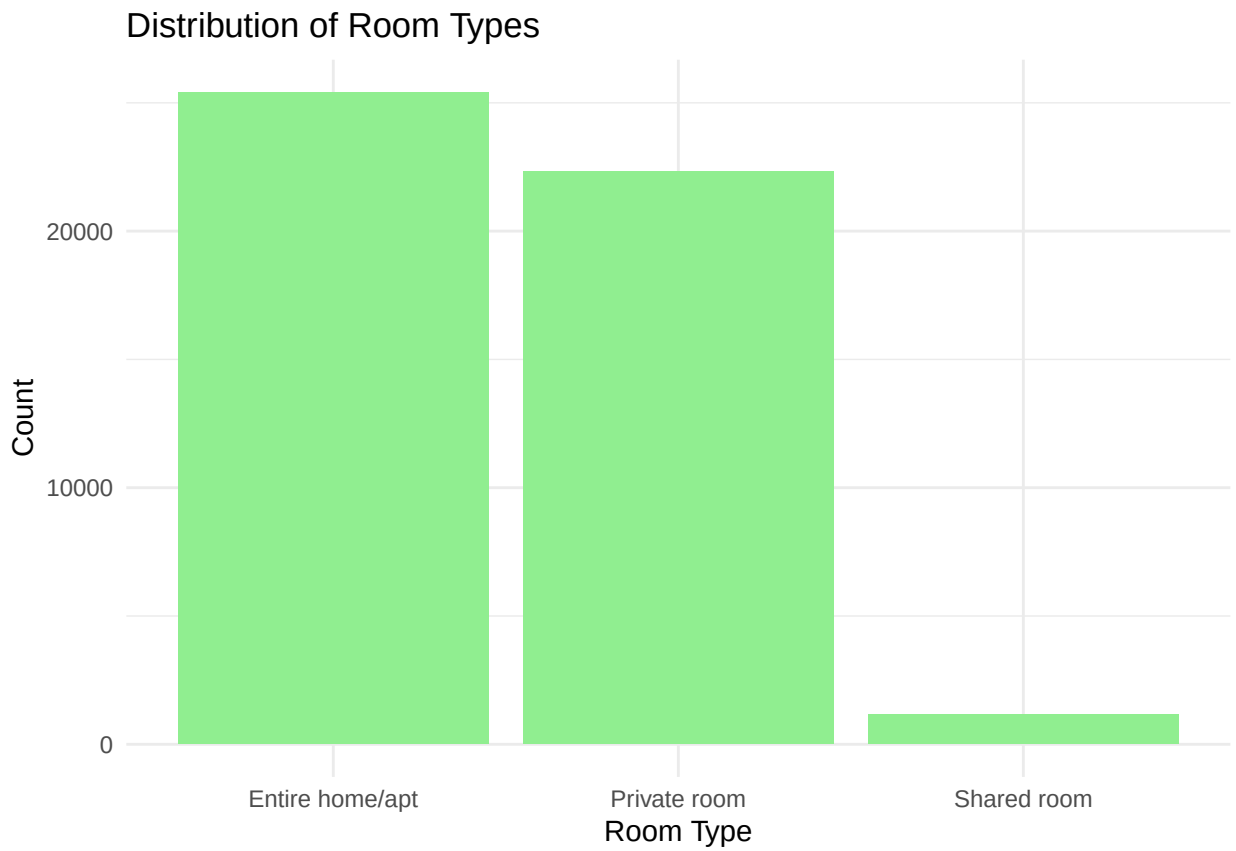
- Private room: 22,326 listings with an average monthly review count of 1.14. Median reviews of 0.400 and standard deviation of 1.717. Private rooms have the highest average monthly review compared to entire homes and shared room.
- Shared room: 1,160 listings with an average monthly review count of 1.07. Median reviews of 0.405 and standard deviation of 1.523.

```
# Box plot of Reviews per Month by Room type
ggplot(data, aes(x = room_type, y = reviews_per_month)) +
  geom_boxplot(fill = "lightblue", color = "darkblue") +
  labs(title = "Reviews per Month by Room Type",
       x = "Room Type", y = "Reviews per Month") +
  theme_minimal()
```



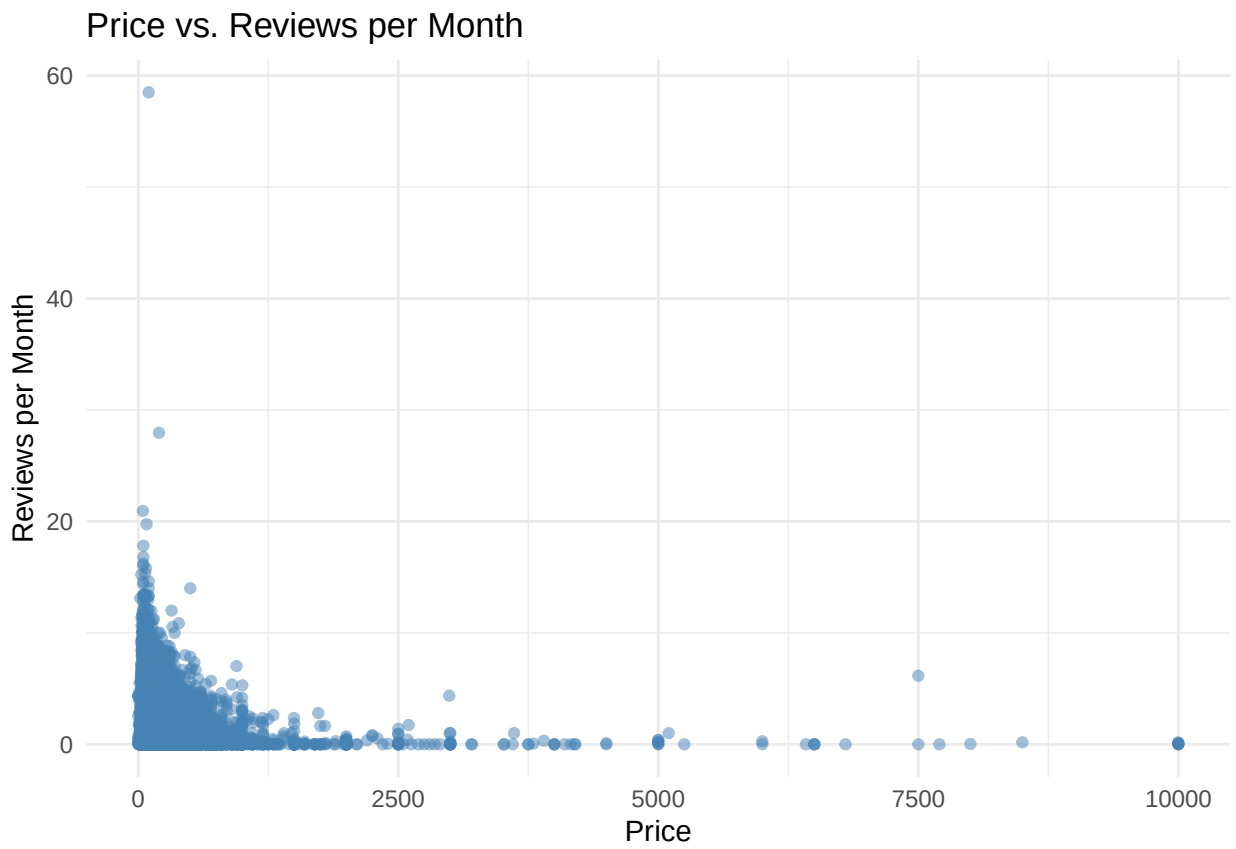
The majority of reviews per month are skewed towards the lower values for all the room types. Moreover, extreme outliers are present, especially for the "private room" type. This indicates a few private rooms recieve far more reviews. Most Airbnb listings, regardless of type, have low monthly review counts.

```
# Distribution of room types
ggplot(data, aes(x = room_type)) +
  geom_bar(fill = "lightgreen") +
  labs(title = "Distribution of Room Types",
       x = "Room Type", y = "Count") +
  theme_minimal()
```



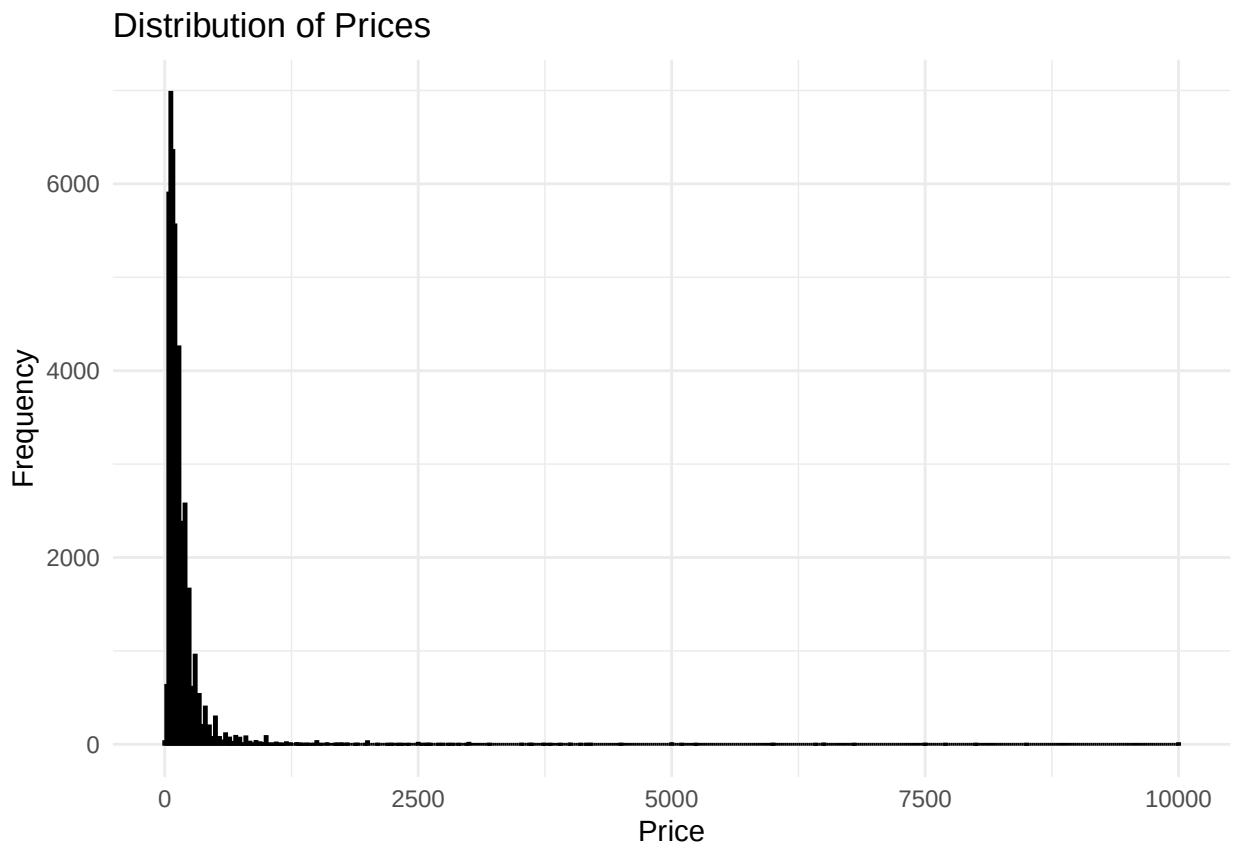
"Entire home/apartment" is the most common room type, followed by "private room," and "shared room" is the least common. This distribution chart indicates that "Entire home/apt" and "Private room" are the dominant room types offered on Airbnb, with shared rooms being a minority.

```
# Price vs. Reviews per Month
ggplot(data, aes(x = price, y = reviews_per_month)) +
  geom_point(alpha = 0.5, color = "steelblue") +
  labs(title = "Price vs. Reviews per Month",
       x = "Price", y = "Reviews per Month") +
  theme_minimal()
```



The relationship between price and reviews is non-linear. There is an inverse relationship betweeen review counts and prices. This suggests affordability encourages more reviews. Higher prices correlate with lower review counts. However, there are outliers that indicate some high-priced properties recieve a higher number of reviews despite this trend. This plot reveals a negative correlation between price and review frequency on Airbnb, highlighting that affordable listings tend to attract more reviews.

```
# Distribution of prices
ggplot(data, aes(x = price)) +
  geom_histogram(binwidth = 20, fill = "purple", color = "black") +
  labs(title = "Distribution of Prices",
       x = "Price", y = "Frequency") +
  theme_minimal()
```



This indicates prices are heavily skewed to the right - most listings are priced at under $1000 with some listings significantly higher. The price skew indicates that hosts predominantly target budget-conscious guests for listings under 1000 dollars, though a small luxury market exists for high-value listings.

Hypothesis 1: ANOVA Test

Since our dataset is finitely large, over 40000 values, due to Central Limit Theorem (CLT), normality can be assumed as the sample size for the dataset is n $>$ 30. The CLT states that, for sufficiently large sample sizes, for n $>$ 30, the sampling distribution of the mean will tend to be normal, regardless of the shape of the underlying data distribution. With over 40,000 observations, the distribution can be assume to be normal.

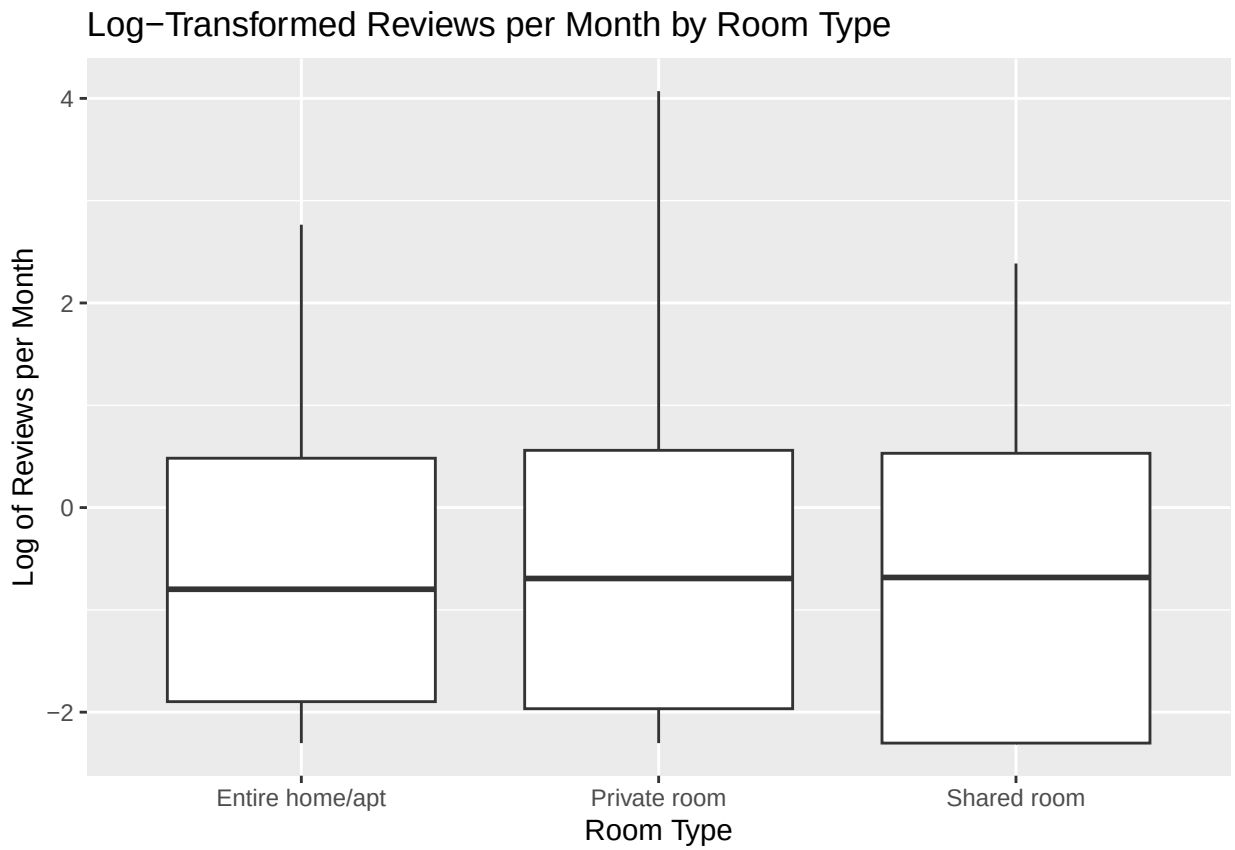
Null Hypothesis (H_0): The type of room (e.g., private room, entire home/apartment) has no significant effect on the number of reviews per month and price.
Alternative Hypothesis (H_1): The type of room has a significant effect on both the number of reviews per month and price

```
# ANOVA test for Reviews per Month by Room Type
anova_reviews <- aov(reviews_per_month ~ room_type, data = data)
summary(anova_reviews)
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## room_type       2    114   57.23   22.45 1.79e-10 ***
## Residuals   48892 124629    2.55
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null hypothesis is that the mean number of reviews per month is the same across all room types (private room, entire hom/apt, shared room). Our alternative hypothesis is that at least one room type has a mean number of reviews per month that is significantly different from the others. The Pr(>F) value of 1.79e-10 is significantly smaller than 0.05, meaning we reject the null hypothesis. This indicates at least one room type has a mean significantly different from the others. The F-statistic of 22.45 also indicates a significant differences in the mean number of reviews per month among the three room types.

```
# Log transformation for reviews_per_month
data <- data %>%
  mutate(log_reviews_per_month = log(reviews_per_month + 0.1))

# Plot the log-transformed data by Room Type
ggplot(data, aes(x = room_type, y = log_reviews_per_month)) +
  geom_boxplot(outlier.colour = "red", outlier.shape = 16, outlier.size = 2) +
  labs(title = "Log-Transformed Reviews per Month by Room Type",
       x = "Room Type", y = "Log of Reviews per Month")
```



Since the raw data was not normally distributed, this log transformation was applied to reviews_per_month to reduce the variance and to approximate normality. Although the mean and median seem close across different room types, there is still a visible difference indicating that there is a relationship between room type and reviews.

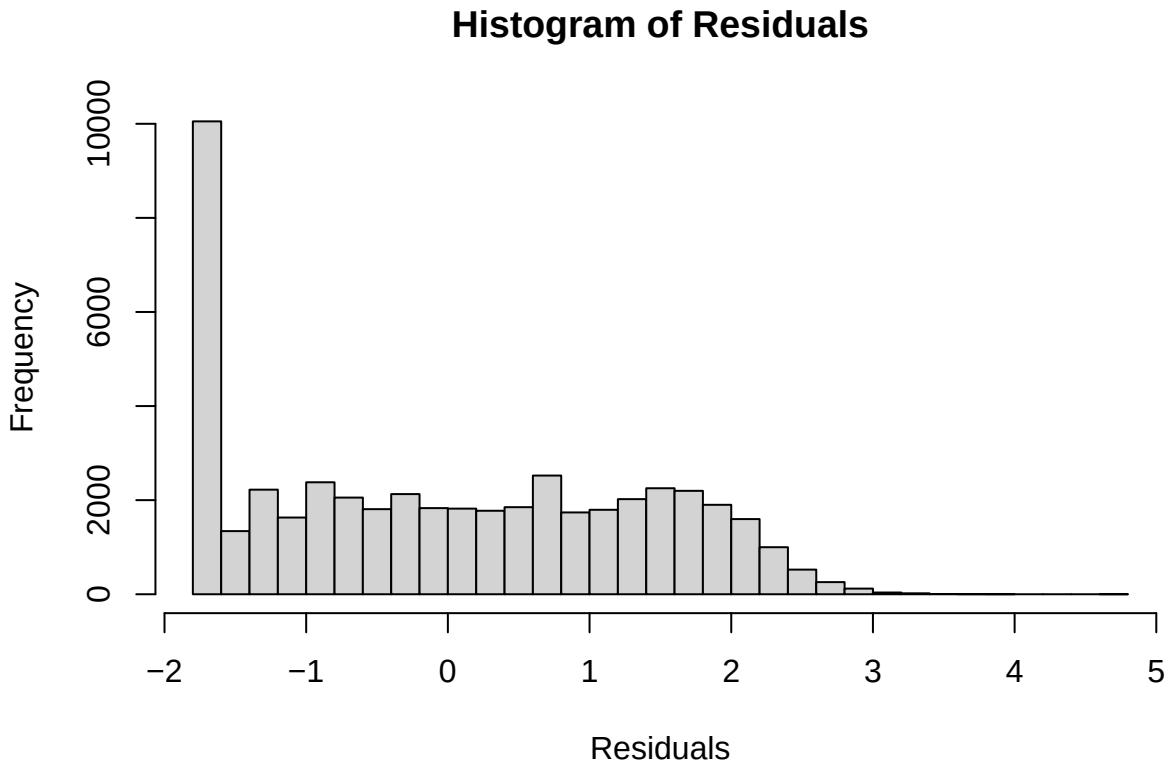
Check for Normality of Residuals

```r
# ANOVA test for Reviews per Month by Room Type
anova_reviews <- aov(log_reviews_per_month ~ room_type, data = data)
summary(anova_reviews)
```

```
##                Df Sum Sq Mean Sq F value  Pr(>F)
## room_type       2     21  10.368   5.895 0.00276 **
## Residuals   48892  85987   1.759
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value is significant $(0.00276) < 0.05$, we can conclude that there is a statistically significant effect of room_type on log_reviews_per_month. This means that at least one of the room types has a different average review rate from the others. However, lets see if the normality assumptions are valid after the log transformation.

```r
# Histogram of residuals for visual inspection
hist(residuals(anova_reviews), main = "Histogram of Residuals", xlab = "Residuals", breaks = 30)
```



This shows the residuals are not perfectly normal. The histogram appears to be skewed to the right, with a very high frequency at the far left end of the histogram, an approximately symmetric central part, and a typical right-skewing pattern towards the right side of the histogram.

```r
# Kolmogorov-Smirnov Test for normality of residuals
ks_test <- ks.test(residuals(anova_reviews),
                   "pnorm", mean = mean(residuals(anova_reviews)),
                   sd = sd(residuals(anova_reviews)))
```

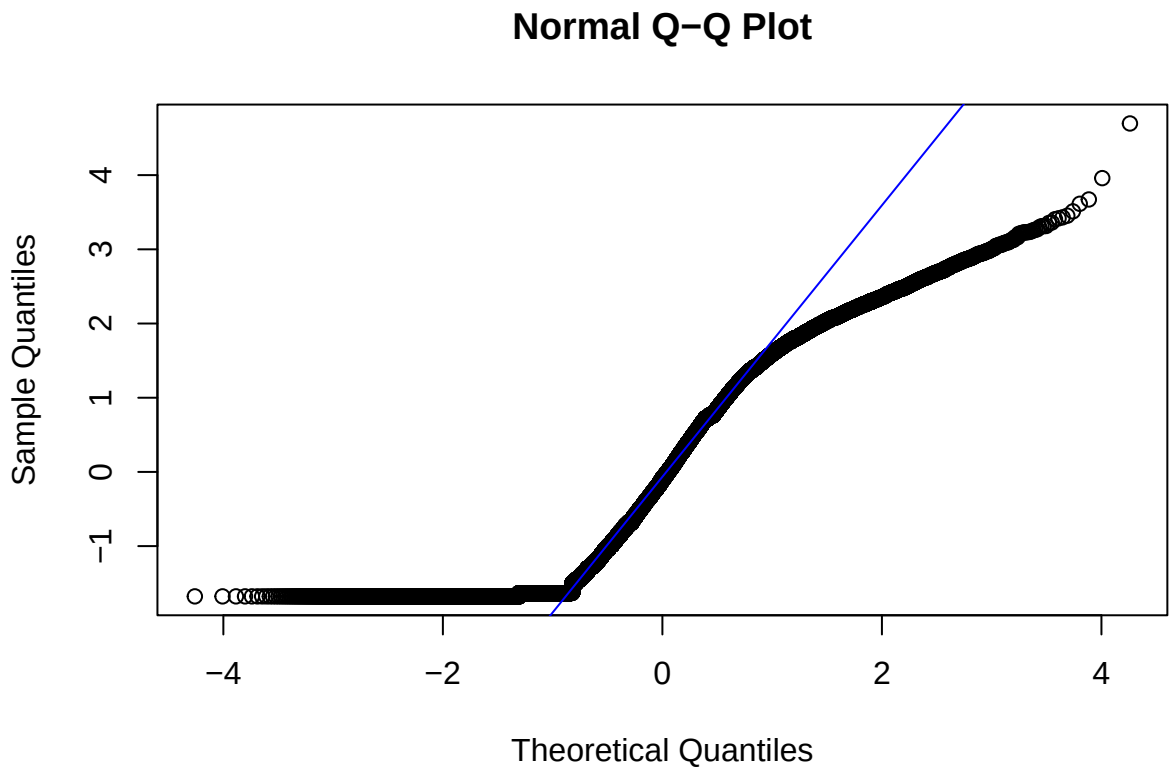
```
## Warning in ks.test.default(residuals(anova_reviews), "pnorm", mean =
## mean(residuals(anova_reviews)), : ties should not be present for the one-sample
## Kolmogorov-Smirnov test
```

```r
print(ks_test)
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  residuals(anova_reviews)
## D = 0.10299, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

The shapiro-wilk test for normality only allows sample size of 5 to 5000, thus, shapiro-wilk is not a viable normality for this Airbnb dataset. Thus, opting to use Kolmogorov-Smirnov Test for normality of residuals. The result test came back with a p-value of <2.2e-16 indicates that residuals deviate from a normal distribution.

```r
qqnorm(residuals(anova_reviews))
qqline(residuals(anova_reviews), col = "blue")
```



Deviations at the tails further confirm residuals are not normally distributed.

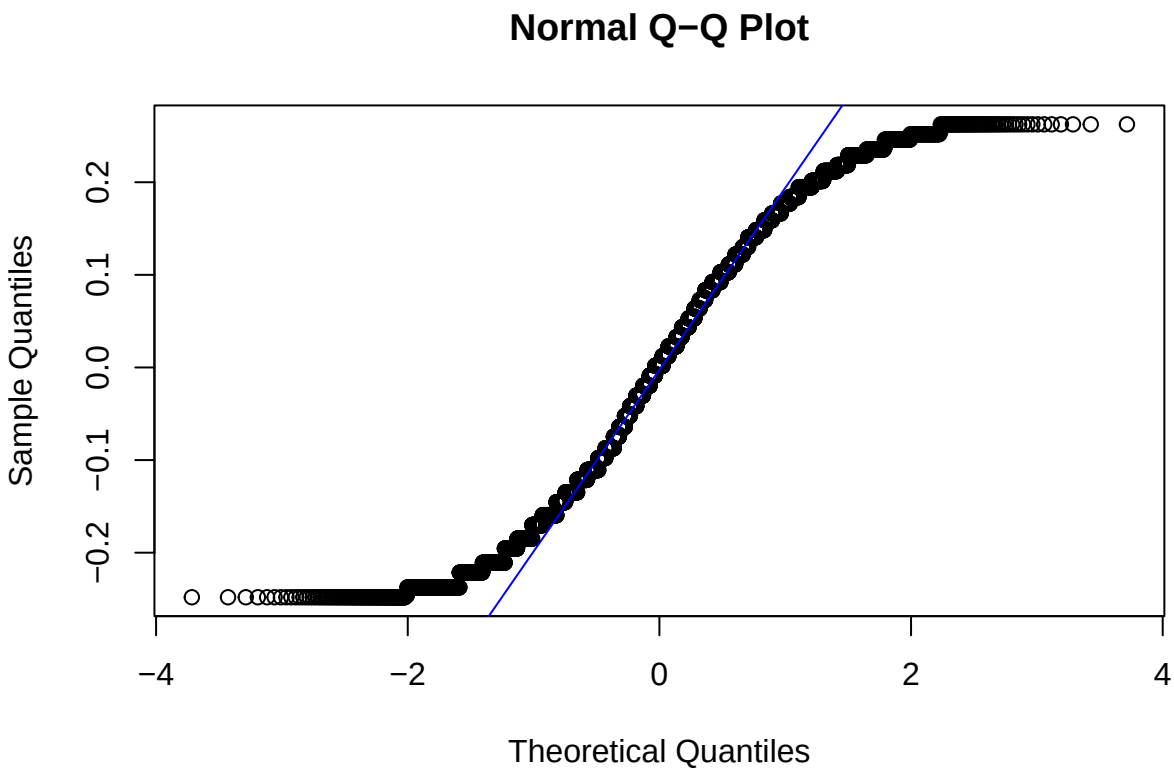
Trimming Data to middle 10% Quantile + ANOVA

```r
# Calculate the 45th and 55th percentiles for reviews_per_month
lower_bound <- quantile(data$reviews_per_month, 0.45, na.rm = TRUE)
upper_bound <- quantile(data$reviews_per_month, 0.55, na.rm = TRUE)

# Filter data to include only the middle 10% (only 45th to 55th percentile)
middle_data <- data %>%
  filter(reviews_per_month >= lower_bound & reviews_per_month <= upper_bound)

# Transformation of the middle quantile
middle_data <- middle_data %>%
  mutate(log_reviews_per_month = log(reviews_per_month + 0.1))

# Fit the ANOVA model on the filtered data
anova_middle <- aov(log_reviews_per_month ~ room_type, data = middle_data)
summary(anova_middle)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## room_type      2   0.14 0.06877   3.071 0.0465 *
## Residuals   4936 110.53 0.02239
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

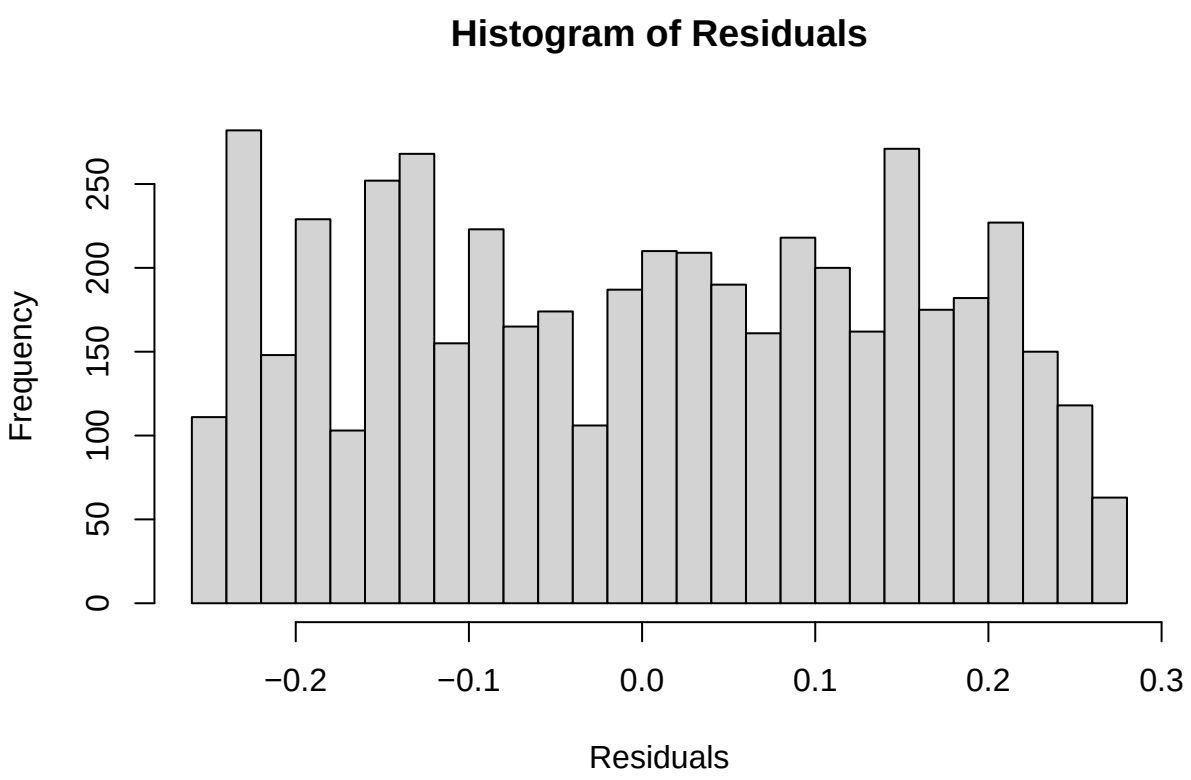
Here, we chose to look at the middle 45% to 55% percent of the data to eliminate outliers and improve normality. The F-statistic of 3.071 and p-value of 0.0465 indicate that after the transformation, room type still significantly affects reviews_per_month, but less strongly, particularly due to the smaller sample size.

```r
# Check assumptions on the middle quantile data
# Q-Q plot for residuals
qqnorm(residuals(anova_middle))
qqline(residuals(anova_middle), col = "blue")
```



The points in the middle appear align closely with the line, while points at the tails show more deviation. This indicates that while the central part of the residuals is approximately normal, the extreme values still show skewness.

```
# Histogram of residuals
hist(residuals(anova_middle), main = "Histogram of Residuals", xlab = "Residuals", breaks = 30)
```



This is approximately symmetric centered at around 0.0. There is less skewness compared to the untrimmed data. This relatively uniform distribution suggests that using ANOVA is useful after the transformation.

```
# Shapiro-Wilk test for normality for the middle 45%-55% quantile
shapiro_test_middle <- shapiro.test(residuals(anova_middle))
print(shapiro_test_middle)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(anova_middle)
## W = 0.95233, p-value < 2.2e-16
```

Indicates that residuals still aren't perfectly normally distributed because of p-value = <2.2e-16, indicating that the residuals are still not perfectly normal.

```
# Levene's test for homogeneity of variances
levene_test_middle <- leveneTest(log_reviews_per_month ~ room_type, data = middle_data)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
print(levene_test_middle)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
```

```
##         Df F value Pr(>F)
## group    2  1.2658 0.2821
##       4936
```

With a p-value of 0.2821, the test indicates homogeneity of variances and indicating that variances across room types are approximately equal within the middle 10% data subset, which supports the use of ANOVA despite not satisfying normality of residuals.

Overall Analysis for Hypothesis 1 with the ANOVA Test: ANOVA shows there is significant association between room type and reviews per month, which remains after the data cleaning plus trimming, although at a weaker strength. The F-statistic and p-values from both ANOVA tests confirm that at least one room type has a mean number of reviews that significantly differs from the rest. Despite having residuals deviate from normality, even after the transformations. Levene's Test supports the use of ANOVA. Overall, the test results supports the alternative hypothesis and reject the null hypothesis at the significance level of $\alpha = 0.05$ as both the p-values (1.79e-10 for the full dataset; 0.0465 for the middle 10% quantile) for the test are less than 0.05. The analysis indicates that room type significantly influences review frequency.

Hypothesis 2: Borough Influence on Reviews per Month

Null Hypothesis (H_0): The neighborhood of the property has NO significant effect on the number of reviews received monthly.
Alternative Hypothesis (H_1)): The neighborhood of the property significantly affects the number of reviews received monthly.

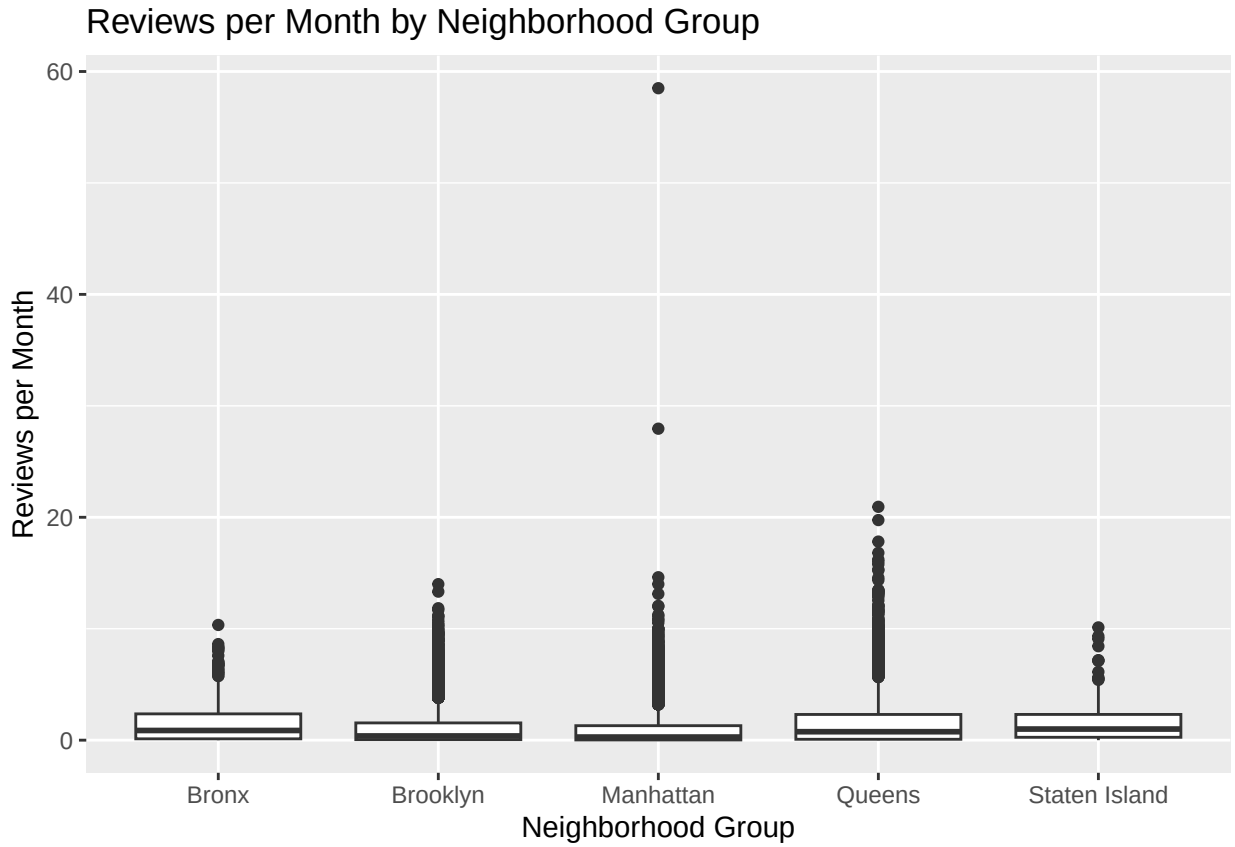
```
# Summary Statistucs of Borough
data %>%
  group_by(neighbourhood_group) %>%
  summarise(
    count = n(),
    avg_reviews_per_month = mean(reviews_per_month, na.rm = TRUE),
    median_reviews_per_month = median(reviews_per_month, na.rm = TRUE)
  )
```

```
## # A tibble: 5 x 4
##   neighbourhood_group count avg_reviews_per_month median_reviews_per_month
##   <chr>               <int>                 <dbl>                     <dbl>
## 1 Bronx                1091                 1.48                       0.87
## 2 Brooklyn            20104                 1.05                       0.38
## 3 Manhattan           21661                 0.977                      0.28
## 4 Queens               5666                 1.57                       0.76
## 5 Staten Island         373                 1.58                       1
```

Hypothesis 2 Analysis : To answer the question of whether or not Neighborhoods (Boroughs) impact review count, the conclusion can be made by looking at the summary statistics, employing a Kruskal-Wallis Test for non-parametric variables that does not assume normality.

From the summmary statistic, the average reviews rates varied throughout the 5 boroughs and concluded that Queens (1.567), Staten Island (1.576) and The Bronx (1.476) have significantly more reviews per month than Manhattan (0.977) and Brooklyn (1.050), however Manhattan and Brooklyn have significantly more reviews all time with 21661 and 20104 count respectively. Manhattan showed the lowest average review rate, indicating potential saturation or less frequent review activity for higher-priced listings. This can be seen in the summary statistic provided.

```r
#  Review per Month by Neighborhood Group
ggplot(data, aes(x = neighbourhood_group, y = reviews_per_month)) +
  geom_boxplot() +
  labs(title = "Reviews per Month by Neighborhood Group", x = "Neighborhood Group", y = "Reviews per Mo
```



From the box plot, it can be seen that the median is relatively close to zero for all neighborhoods, indicating that a majority of listings receive a low number of reviews per month. Queens appears to have a slightly higher range of reviews within its IQR compared to other neighborhoods, suggesting it may attract slightly more consistent engagement with users/reviews. Each neighborhood group displays great number of outliers. These outliers indicate listings with a higher-than-average number of reviews per month. Queens and Brooklyn have more outliers than other neighborhoods, showing that some listings in these areas receive significantly more reviews compared to the others in this sample. This might reflect areas with popular listings or high demand within these neighborhoods. Manhattan also very has few notable outliers, but fewer compared to Queens and Brooklyn, which could be due to competition or higher pricing and thus limiting frequent reviews, but there are few ones that have the most number of reviews among the boroughs.

It is important to note that Manhattan has a wide range of values, including a significant amount of low average reviews per month properties, indicating that some listings can receive a lot of engagement and reviews while others can receive little to none. For the overall distribution, the review counts are highly skewed towards the lower end, with most listings across all neighborhoods having relatively low monthly review counts. The presence of many low or zero-review listings and a few high-review outliers suggests that only a small fraction of properties in each neighborhood receive reviews and engagement.

```r
# Kruskal-Wallis test for Reviews per Month by Neighborhood Group
kruskal_test <- kruskal.test(reviews_per_month ~ neighbourhood_group, data = data)
kruskal_test
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  reviews_per_month by neighbourhood_group
## Kruskal-Wallis chi-squared = 587.12, df = 4, p-value < 2.2e-16
```

Earlier, it was shown with normality assumptions, the dataset isn't the best for parametric tests, so the Kruskal-Wallis test which is non-parametric can be used to determine if there are statistically significant differences between the distributions of a continuous variable across multipe boroughs. Since the Kruskal-Wallis test is non-parametric, it is based on ranks rather than actual values unlike ANOVA (which needs normal distribution).

With the results of the Kruskal-Wallis Test, there was significant statistical difference between the boroughs. Since the p-value is much smaller than the significance level $\alpha = 0.05$, we reject the null hypothesis of the Kruskal-Wallis test. This means there is a statistically significant difference in the distribution of reviews_per_month across the different neighbourhood_group categories. This result suggests that the frequency of reviews per month is not uniform across different neighborhoods. Some neighborhoods tend to have significantly different review frequencies than others, which aligns with the patterns observed in the earlier boxplot analysis. Thus, it can be concluded that the property in certain New York City neighborhoods do in fact, receive more reviews on average monthly.