AMS 317 Linear Regression Project

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

Due Date: November 10, 2024

Hypothesis 1: Does **room type** (e.g., private room vs. entire home/apartment) influence the number of **reviews per month**? Hypothesis 2: Do properties in certain **neighborhoods** receive significantly **more reviews** than others?

```
# Load data and check for missing values
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)

data <- read_csv("/Users/hyl/Downloads/AB_NYC_2019.csv")
```

```
## Rows: 48895 Columns: 16

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (6): name, host_name, neighbourhood_group, neighbourhood, room_type, la...
## dbl (10): id, host_id, latitude, longitude, price, minimum_nights, number_of...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# 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
```

```
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.

```
# 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
```

```
# 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_m


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

```
# 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)
  )
```

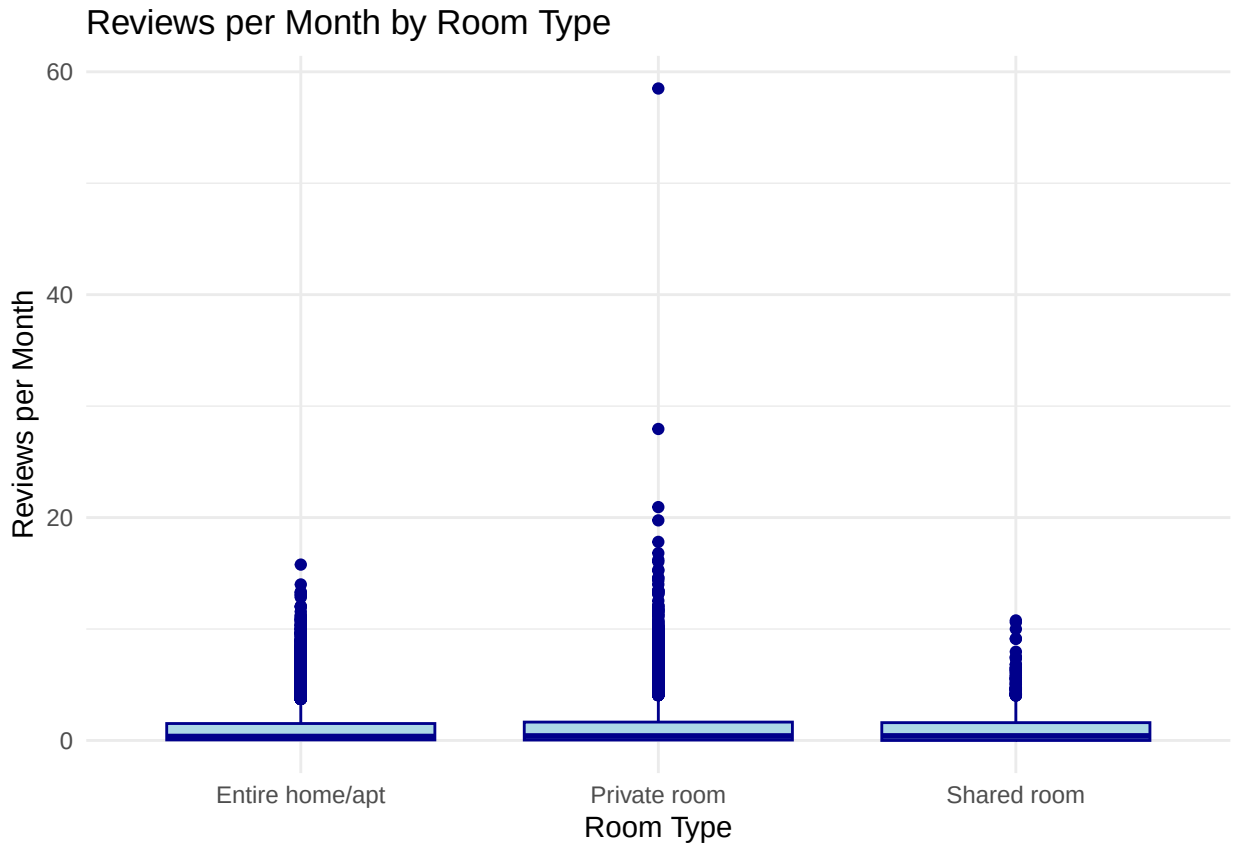
```
## # A tibble: 3 x 4
##   room_type       count avg_price avg_reviews_per_month
##   <chr>           <int>     <dbl>                 <dbl>
## 1 Entire home/apt 25409     212.                   1.05
## 2 Private room    22326      89.8                  1.14
## 3 Shared room      1160      70.1   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

```
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
```

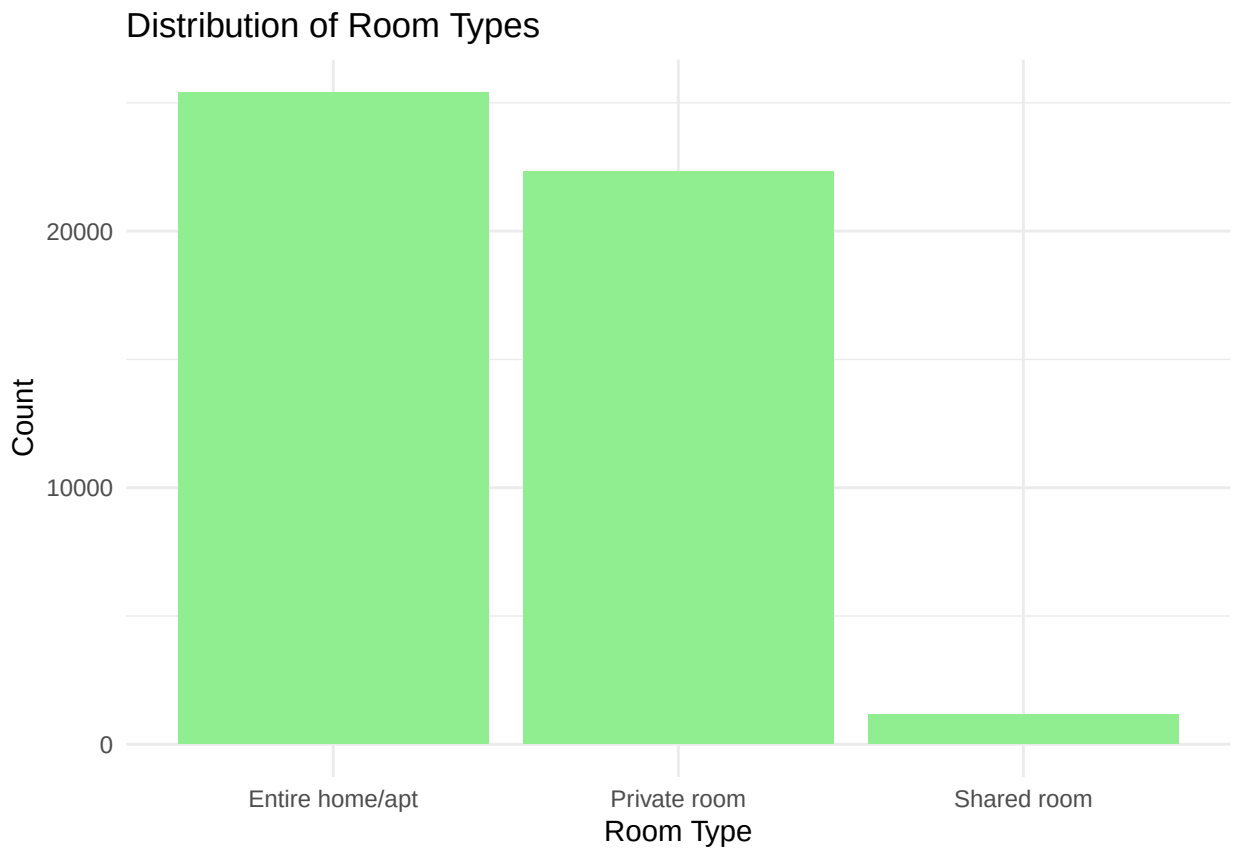
```
## # A tibble: 3 x 4
##   room_type       mean_reviews median_reviews sd_reviews
##   <chr>                  <dbl>          <dbl>      <dbl>
```

```
## 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
```

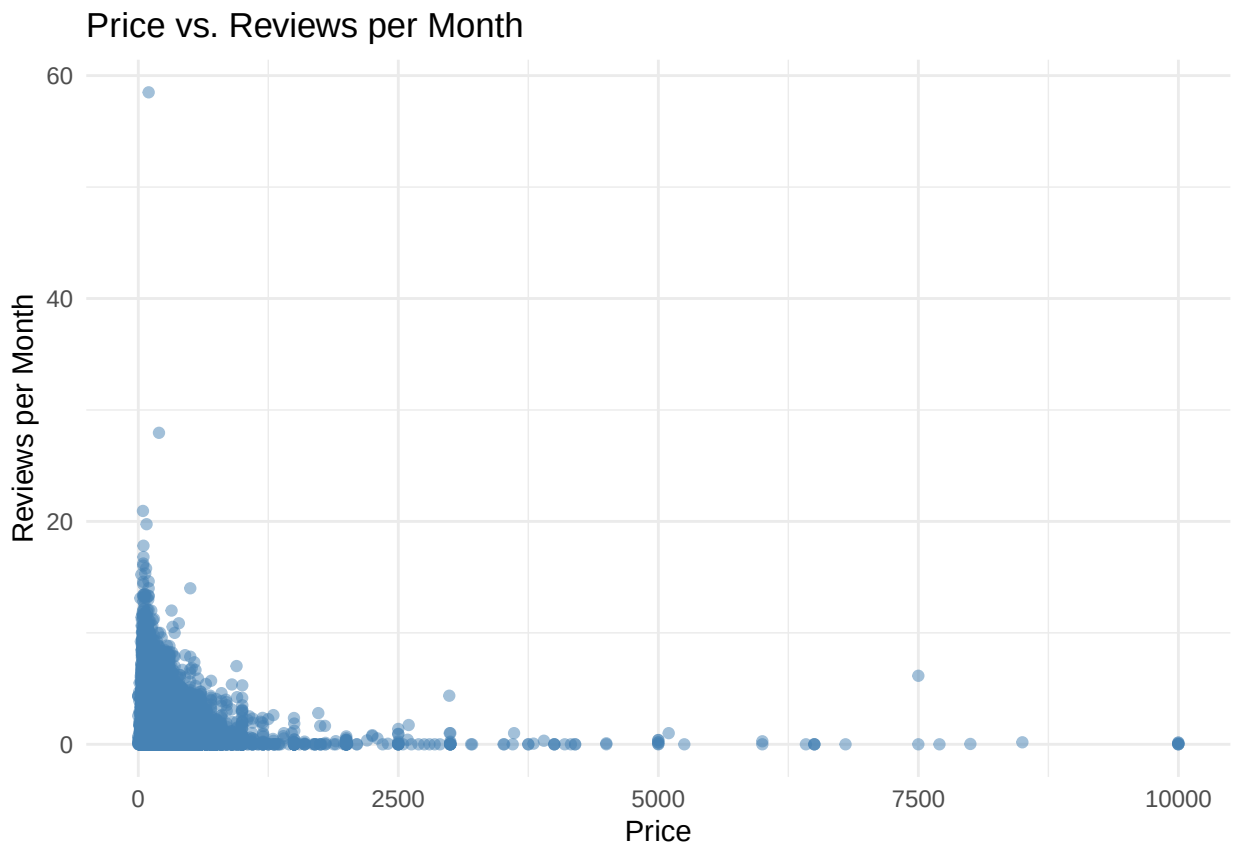
```r
# Box plot of Reviews per Month by Room type
ggplot(data, aes(x = room_type, y = reviews_per_month)) +
  geom_boxplot(fill = "lightblue", color = "darkblue") +
  labs(title = "Reviews per Month by Room Type",
       x = "Room Type", y = "Reviews per Month") +
  theme_minimal()
```



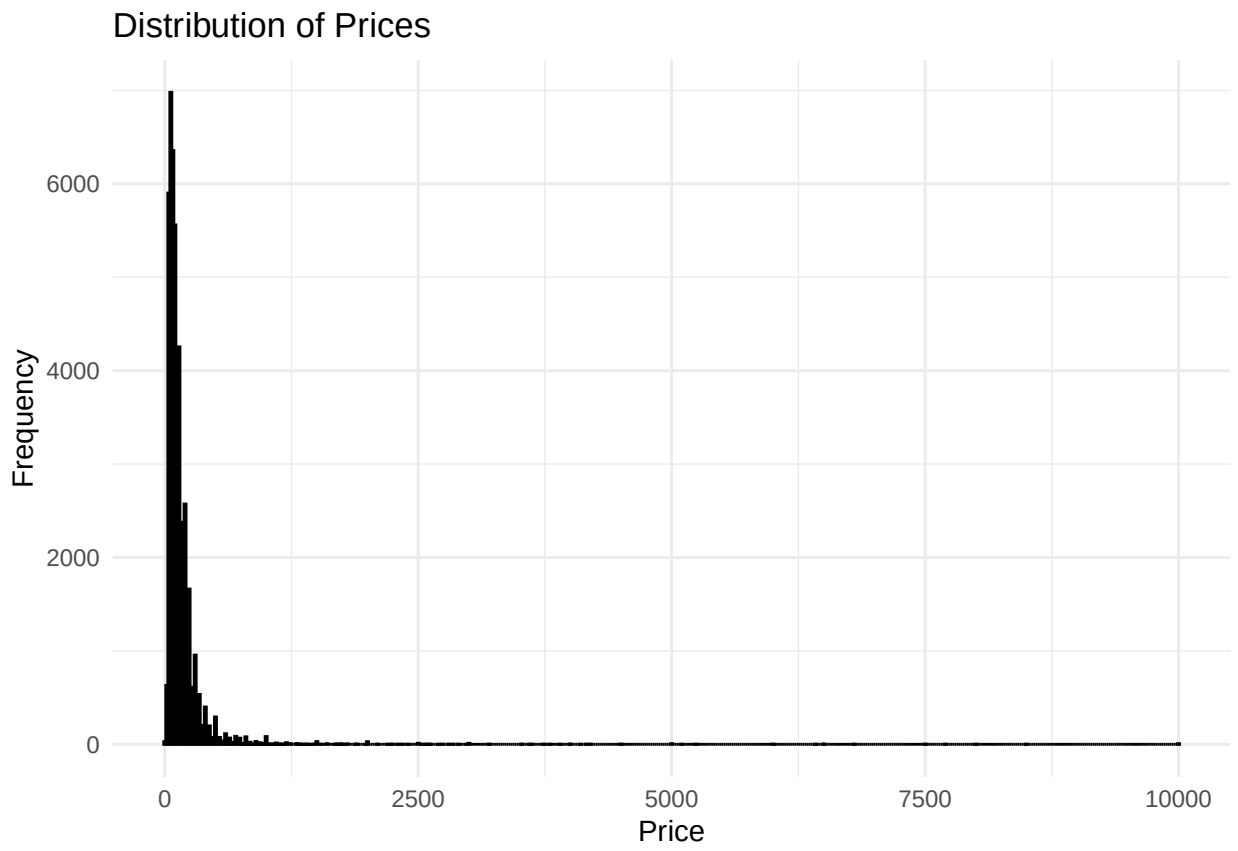

```r
# Distribution of room types
ggplot(data, aes(x = room_type)) +
  geom_bar(fill = "lightgreen") +
  labs(title = "Distribution of Room Types",
       x = "Room Type", y = "Count") +
  theme_minimal()
```




```
# 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()
```




```r
# 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()
```



Hypothesis 1: ANOVA Test

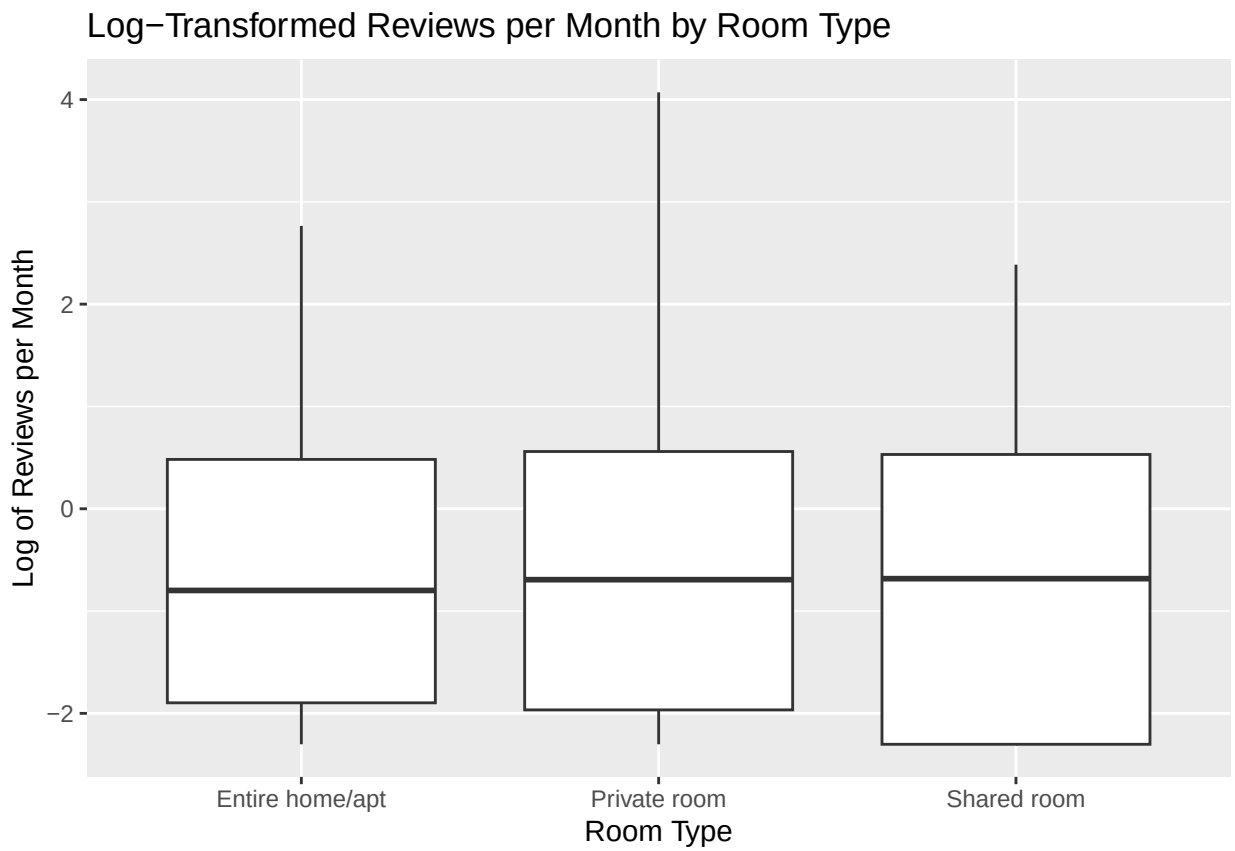
Since our dataset is finitely large, over 40000 values, we can assume normality in our data. Thus, performing anova suffices.

```
# 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
```

```
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")
```



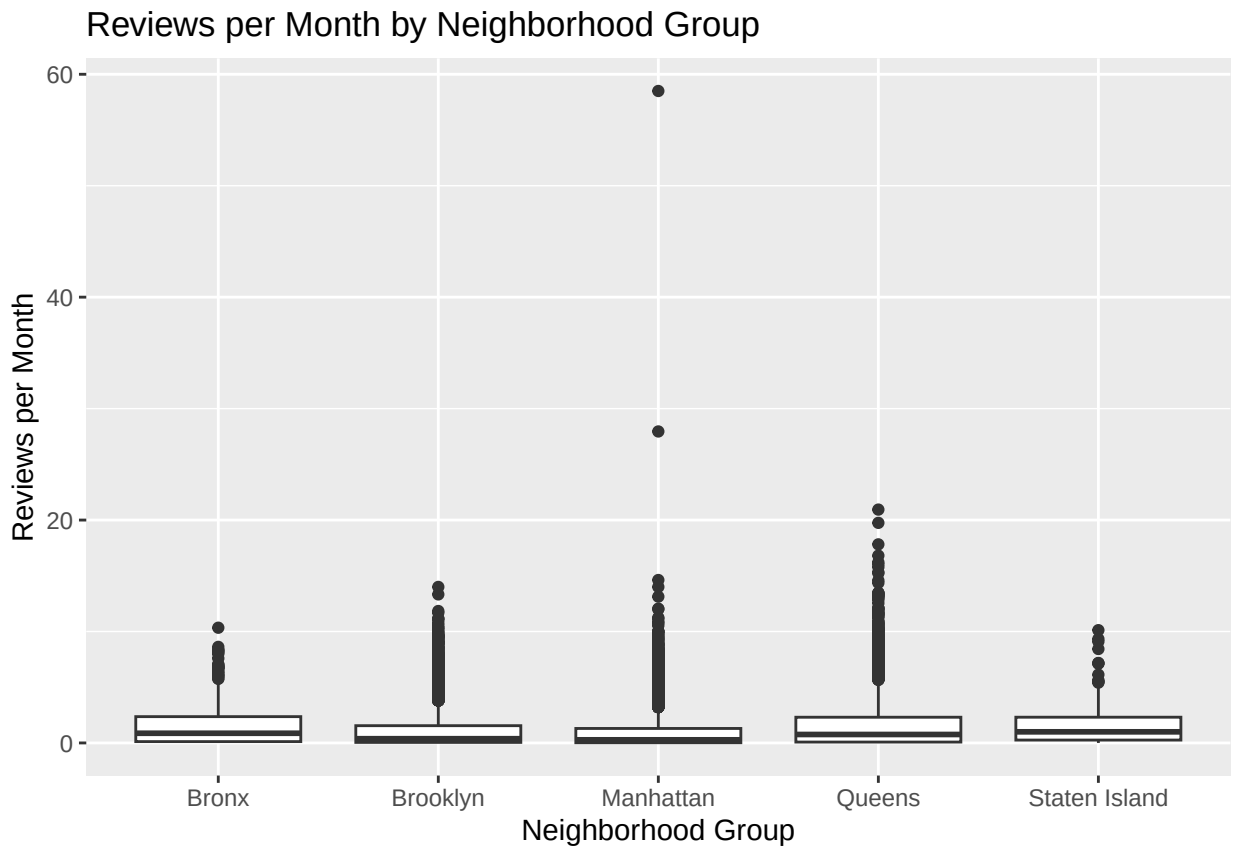
Hypothesis 2: Borough Influence on Reviews per Month

```
# 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
```

```
#  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
```




```
# 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
```