AMS 317 Linear Regression Project

Group 3:

Due Date: November 10, 2024

Read Data

```
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
```

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

```
## Rows: 48895 Columns: 16

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (5): name, host_name, neighbourhood_group, neighbourhood, room_type
## dbl  (10): id, host_id, latitude, longitude, price, minimum_nights, number_o...
## date  (1): last_review
##
## 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.
```

```
# Data Inspection
str(data)
```

```
## spc_tbl_ [48,895 x 16] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ id                            : num [1:48895] 2539 2595 3647 3831 5022 ...
##  $ name                          : chr [1:48895] "Clean & quiet apt home by the park" "Skylit Midtow
##  $ host_id                       : num [1:48895] 2787 2845 4632 4869 7192 ...
##  $ host_name                     : chr [1:48895] "John" "Jennifer" "Elisabeth" "LisaRoxanne" ...
##  $ neighbourhood_group           : chr [1:48895] "Brooklyn" "Manhattan" "Manhattan" "Brooklyn" ...
##  $ neighbourhood                 : chr [1:48895] "Kensington" "Midtown" "Harlem" "Clinton Hill" ...
##  $ latitude                      : num [1:48895] 40.6 40.8 40.8 40.7 40.8 ...
##  $ longitude                     : num [1:48895] -74 -74 -73.9 -74 -73.9 ...
##  $ room_type                     : chr [1:48895] "Private room" "Entire home/apt" "Private room" "En
##  $ price                         : num [1:48895] 149 225 150 89 80 200 60 79 79 150 ...
##  $ minimum_nights                : num [1:48895] 1 1 3 1 10 3 45 2 2 1 ...
##  $ number_of_reviews             : num [1:48895] 9 45 0 270 9 74 49 430 118 160 ...
##  $ last_review                   : Date[1:48895], format: "2018-10-19" "2019-05-21" ...
##  $ reviews_per_month             : num [1:48895] 0.21 0.38 NA 4.64 0.1 0.59 0.4 3.47 0.99 1.33 ...
##  $ calculated_host_listings_count: num [1:48895] 6 2 1 1 1 1 1 1 1 4 ...
##  $ availability_365              : num [1:48895] 365 355 365 194 0 129 0 220 0 188 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   id = col_double(),
##   ..   name = col_character(),
##   ..   host_id = col_double(),
##   ..   host_name = col_character(),
##   ..   neighbourhood_group = col_character(),
##   ..   neighbourhood = col_character(),
##   ..   latitude = col_double(),
##   ..   longitude = col_double(),
##   ..   room_type = col_character(),
##   ..   price = col_double(),
##   ..   minimum_nights = col_double(),
##   ..   number_of_reviews = col_double(),
##   ..   last_review = col_date(format = ""),
##   ..   reviews_per_month = col_double(),
##   ..   calculated_host_listings_count = col_double(),
##   ..   availability_365 = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
head(data)
```

```
## # A tibble: 6 x 16
##      id name        host_id host_name neighbourhood_group neighbourhood latitude
##   <dbl> <chr>         <dbl> <chr>     <chr>               <chr>            <dbl>
## 1  2539 Clean & qu~    2787 John      Brooklyn            Kensington        40.6
## 2  2595 Skylit Mid~    2845 Jennifer  Manhattan           Midtown           40.8
## 3  3647 THE VILLAG~    4632 Elisabeth Manhattan           Harlem            40.8
## 4  3831 Cozy Entir~    4869 LisaRoxa~ Brooklyn            Clinton Hill      40.7
## 5  5022 Entire Apt~    7192 Laura     Manhattan           East Harlem       40.8
## 6  5099 Large Cozy~    7322 Chris     Manhattan           Murray Hill       40.7
## # i 9 more variables: longitude <dbl>, room_type <chr>, price <dbl>,
## #   minimum_nights <dbl>, number_of_reviews <dbl>, last_review <date>,
## #   reviews_per_month <dbl>, calculated_host_listings_count <dbl>,
## #   availability_365 <dbl>
```

```
summary(data)
```

```
##        id                name              host_id          host_name
##  Min.   :    2539   Length:48895       Min.   :     2438   Length:48895
##  1st Qu.: 9471945   Class :character   1st Qu.:  7822033   Class :character
##  Median :19677284   Mode  :character   Median : 30793816   Mode  :character
##  Mean   :19017143                      Mean   : 67620011
##  3rd Qu.:29152178                      3rd Qu.:107434423
##  Max.   :36487245                      Max.   :274321313
##
##  neighbourhood_group neighbourhood         latitude       longitude
##  Length:48895        Length:48895       Min.   :40.50   Min.   :-74.24
##  Class :character    Class :character   1st Qu.:40.69   1st Qu.:-73.98
##  Mode  :character    Mode  :character   Median :40.72   Median :-73.96
##                                         Mean   :40.73   Mean   :-73.95
##                                         3rd Qu.:40.76   3rd Qu.:-73.94
##                                         Max.   :40.91   Max.   :-73.71
##
##   room_type             price         minimum_nights    number_of_reviews
##  Length:48895       Min.   :    0.0   Min.   :   1.00   Min.   :  0.00
##  Class :character   1st Qu.:   69.0   1st Qu.:   1.00   1st Qu.:  1.00
##  Mode  :character   Median :  106.0   Median :   3.00   Median :  5.00
##                     Mean   :  152.7   Mean   :   7.03   Mean   : 23.27
##                     3rd Qu.:  175.0   3rd Qu.:   5.00   3rd Qu.: 24.00
##                     Max.   :10000.0   Max.   :1250.00   Max.   :629.00
##
##   last_review         reviews_per_month calculated_host_listings_count
##  Min.   :2011-03-28   Min.   : 0.010    Min.   :  1.000
##  1st Qu.:2018-07-08   1st Qu.: 0.190    1st Qu.:  1.000
##  Median :2019-05-19   Median : 0.720    Median :  1.000
##  Mean   :2018-10-04   Mean   : 1.373    Mean   :  7.144
##  3rd Qu.:2019-06-23   3rd Qu.: 2.020    3rd Qu.:  2.000
##  Max.   :2019-07-08   Max.   :58.500    Max.   :327.000
##  NA’s   :10052        NA’s   :10052
##  availability_365
##  Min.   :  0.0
##  1st Qu.:  0.0
##  Median : 45.0
##  Mean   :112.8
##  3rd Qu.:227.0
##  Max.   :365.0
##
```

```r
# Check for NA
data <- data %>%
  filter(!is.na(reviews_per_month), !is.na(price), !is.na(room_type))

# Add a column to show host activity level: single-listing or multiple-listing hosts
data <- data %>%
  group_by(host_id) %>%
  mutate(host_activity = ifelse(n() == 1, "single", "multiple")) %>%
  ungroup()
```

Question 1: Test the Influence of Room on Reviews Per Month

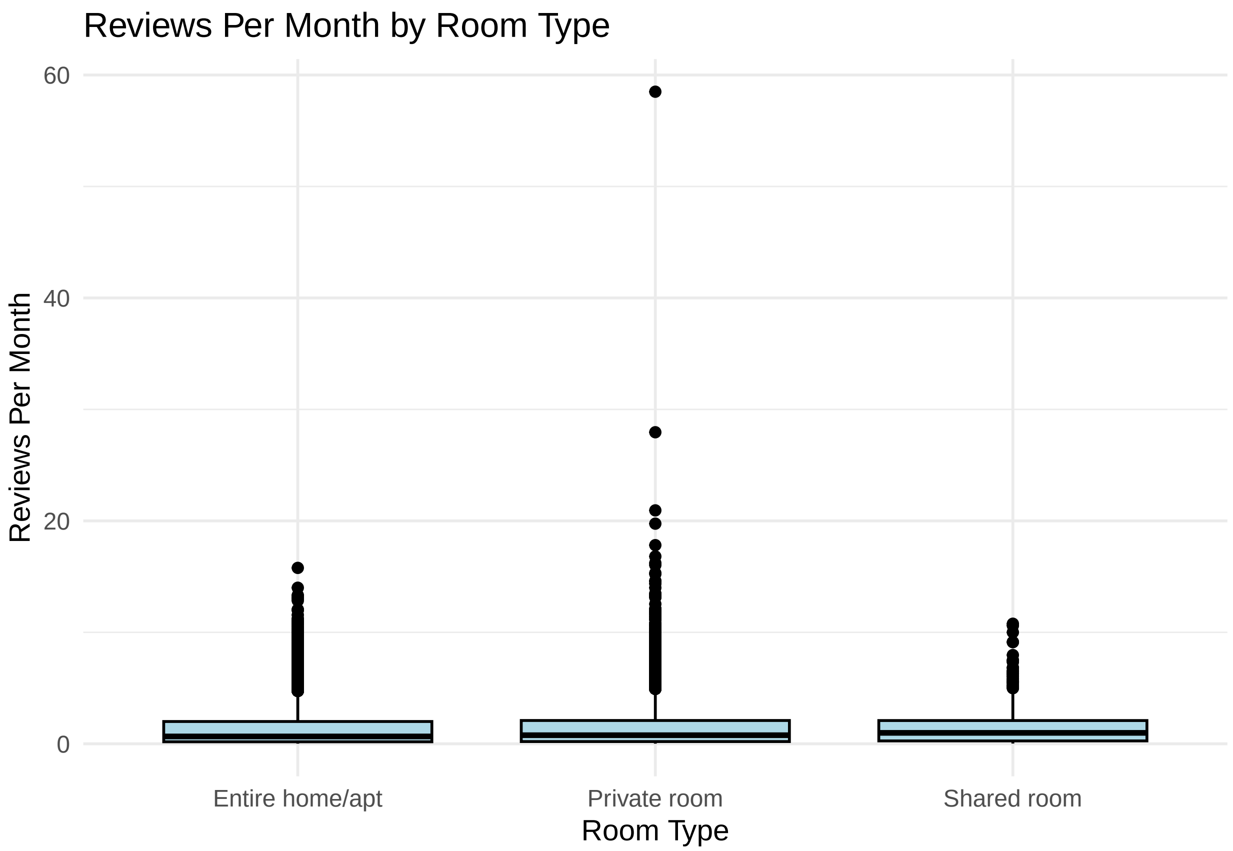
```
anova_result <- aov(reviews_per_month ~ room_type, data = data)
summary(anova_result)
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)    
## room_type       2    190   95.03   33.71 2.36e-15 ***
## Residuals   38840 109495    2.82                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With this p-value: we reject the null hypothesis -> there is a significant difference in the reviews per month across different room types.

```
# Boxplot for Room Type vs Reviews Per Month
library(ggplot2)

ggplot(data, aes(x = room_type, y = reviews_per_month)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(title = "Reviews Per Month by Room Type",
       x = "Room Type",
       y = "Reviews Per Month") +
  theme_minimal()
```

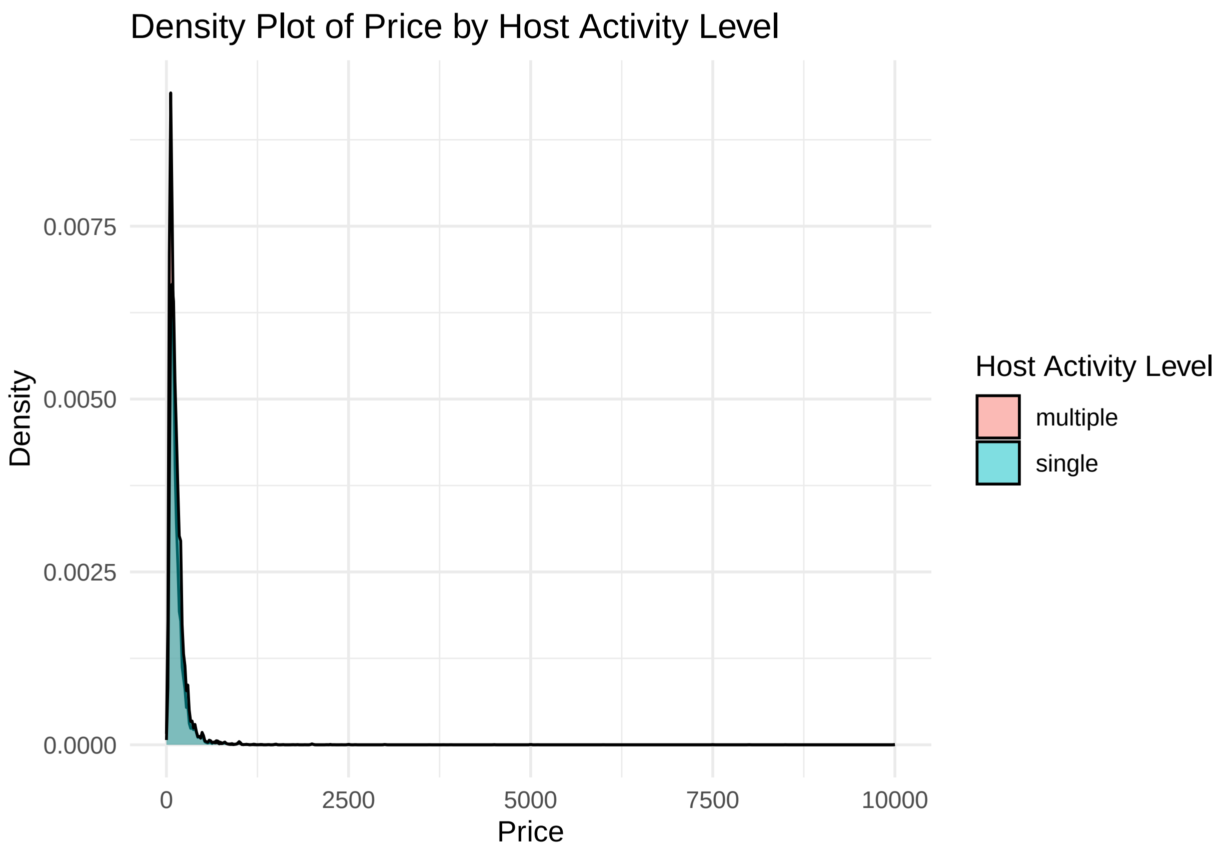


Question 2: Test for Price Differences between Host Activity Level

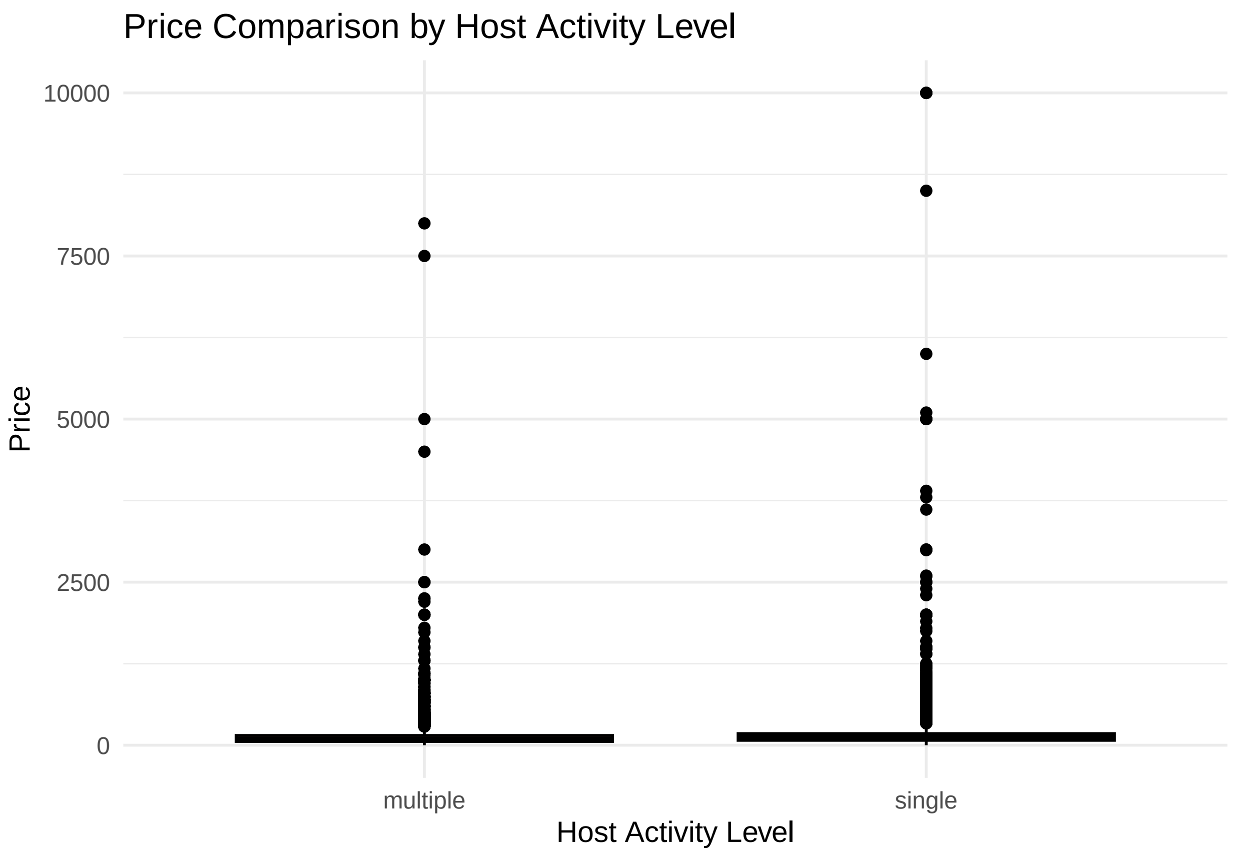
```
t_test_result <- t.test(price ~ host_activity, data = data, var.equal = TRUE)
t_test_result
```

```
##
##  Two Sample t-test
##
## data:  price by host_activity
## t = -11.351, df = 38841, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group multiple and group single is not equal
## 95 percent confidence interval:
##  -28.20033 -19.89558
## sample estimates:
## mean in group multiple   mean in group single
##               126.2528               150.3007
```

```
# Plots for Question 2
# Density plot to show distribution
ggplot(data, aes(x = price, fill = host_activity)) +
  geom_density(alpha = 0.5) +
  labs(title = "Density Plot of Price by Host Activity Level",
       x = "Price",
       y = "Density",
       fill = "Host Activity Level") +
  theme_minimal()
```



```r
# Boxplot
ggplot(data, aes(x = host_activity, y = price)) +
  geom_boxplot(fill = "lightcoral", color = "black") +
  labs(title = "Price Comparison by Host Activity Level",
       x = "Host Activity Level",
       y = "Price") +
  theme_minimal()
```


Price Comparison by Host Activity Level
10000
7500
5000
2500
0
Price
multiple
single
Host Activity Level