Data Visualization Project

Rezaul Karim Mamun, Abdullah All Mamun, Amin Suaad

7/21/2021

BEUTH HOCHSCHULE
FÜR TECHNIK
BERLIN

University of Applied Sciences

CHAPTER 1 : Introduction

CHAPTER 2 : Corelation of variables

CHAPTER 3 : Location and Room Type

CHAPTER 4 : Location And Price

CHAPTER 5 : Popularity of Room based on Price

CHAPTER 6 : Final Analysis

CHAPTER 7 : Summary

Library used in this Project:

ggplot, ggplot2, psych, corrplot, magrittr, dplyr

CHAPTER 1: Introduction

Name of The Dataset: New York City Airbnb Open Data

INTRODUCTION

We are a team of three - Rezaul Karim Mamun, Abdullah All Mamun, Amin Suaad. We present here our visualizations, interactive plots and lots of other interesting insights of the Airbnb data. The Data was collected from Kaggle. In this project we present to you the visualizations of New York Airbnb data.

Airbnb is an American company which operates an online marketplace and hospitality service for people to lease or rent short-term lodging including holiday cottages, apartments, homestays, hostel beds, or hotel rooms, to participate in or facilitate experiences related to tourism such as walking tours, and to make reservations at restaurants. The company does not own any real estate or conduct tours; it is a broker which receives percentage service fees in conjunction with every booking. Like all hospitality services, Airbnb is an example of collaborative consumption and sharing. The company has over 4 million lodging listings in 65,000 cities and 191 countries and has facilitated over 260 million check-ins.

Features of the Dataset:

1. id: The id assigned to each airbnb to identify them uniquely.
2. name: The name assigned to each airbnb.
3. host_id: The id assigned to each host to identify them uniquely.
4. host_name: The name assigned to each host.
5. neighbourhood_group: The 5 boroughs that New York City is divided into: Manhattan, Queens, Brooklyn, Staten Island and Bronx.
6. neighbourhood: The neighborhood where the airbnb is located within the boroughs.
7. latitude: The latitude of the location where the airbnb is situated.
8. longitude: The longitude of the location where the airbnb is situated.
9. room_type: The type of airbnb which is divided into two: Entire home/Apartment, Private room and Shared Room.
10. price: The rent of the airbnb per night.
11. minimum_nights: The minimum number of nights the airbnb can be rented for.
12. number_of_reviews: Total number of reviews posted by customers.
13. last_review: Date of the last review posted by a customer.
14. reviews_per_month: Monthly total of reviews posted by customers.
15. calculated_host_listings_count: Number of total listings by a host.
16. availability_365: Yearly number of days the airbnb is available for rent.

Now, We will try to to find out some basic findings of the dataset and also some basic graphs

```
nycData <- read.csv("AB_NYC_2019.csv") ## Loading the data in R
summary(nycData)  ## statistical summary of the data set
```

```
##        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
##  Length:48895       Min.   : 0.010    Min.   :  1.000               
##  Class :character   1st Qu.: 0.190    1st Qu.:  1.000               
##  Mode  :character   Median : 0.720    Median :  1.000               
##                     Mean   : 1.373    Mean   :  7.144               
##                     3rd Qu.: 2.020    3rd Qu.:  2.000               
##                     Max.   :58.500    Max.   :327.000               
##                     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   
##                  
```

```
head(nycData)
```

```
##     id                                             name host_id   host_name
## 1 2539               Clean & quiet apt home by the park    2787        John
## 2 2595                            Skylit Midtown Castle    2845    Jennifer
## 3 3647              THE VILLAGE OF HARLEM....NEW YORK !    4632   Elisabeth
## 4 3831                  Cozy Entire Floor of Brownstone    4869 LisaRoxanne
## 5 5022 Entire Apt: Spacious Studio/Loft by central park    7192       Laura
## 6 5099        Large Cozy 1 BR Apartment In Midtown East    7322       Chris
##   neighbourhood_group neighbourhood latitude longitude       room_type price
## 1            Brooklyn    Kensington 40.64749 -73.97237    Private room   149
## 2           Manhattan       Midtown 40.75362 -73.98377 Entire home/apt   225
## 3           Manhattan        Harlem 40.80902 -73.94190    Private room   150
## 4            Brooklyn  Clinton Hill 40.68514 -73.95976 Entire home/apt    89
## 5           Manhattan   East Harlem 40.79851 -73.94399 Entire home/apt    80
## 6           Manhattan   Murray Hill 40.74767 -73.97500 Entire home/apt   200
##   minimum_nights number_of_reviews last_review reviews_per_month
## 1              1                 9  2018-10-19              0.21
## 2              1                45  2019-05-21              0.38
## 3              3                 0                            NA
## 4              1               270  2019-07-05              4.64
## 5             10                 9  2018-11-19              0.10
## 6              3                74  2019-06-22              0.59
##   calculated_host_listings_count availability_365
## 1                              6              365
## 2                              2              355
## 3                              1              365
## 4                              1              194
## 5                              1                0
## 6                              1              129
```

Creating descriptive statistics for each variable

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.0.5
```

```
describe(nycData)  ## Description of variables
```

```
##                                vars     n        mean          sd      median
## id                                1 48895 19017143.24 10983108.39 19677284.00
## name*                             2 48895    23962.22    13819.31    23931.00
## host_id                           3 48895 67620010.65 78610967.03 30793816.00
## host_name*                        4 48895     5459.08     3233.51     5329.00
## neighbourhood_group*              5 48895        2.68        0.74        3.00
## neighbourhood*                    6 48895      108.11       68.74       95.00
## latitude                          7 48895       40.73        0.05       40.72
## longitude                         8 48895      -73.95        0.05      -73.96
## room_type*                        9 48895        1.50        0.55        1.00
## price                            10 48895      152.72      240.15      106.00
## minimum_nights                   11 48895        7.03       20.51        3.00
## number_of_reviews                12 48895       23.27       44.55        5.00
## last_review*                     13 48895     1185.55      700.47     1579.00
## reviews_per_month                14 38843        1.37        1.68        0.72
## calculated_host_listings_count   15 48895        7.14       32.95        1.00
## availability_365                 16 48895      112.78      131.62       45.00
##                                    trimmed         mad     min          max
## id                             19188061.30 14689959.59 2539.00  36487245.00
## name*                             23961.15    17717.07    1.00     47906.00
## host_id                        54170438.55 40836605.41 2438.00 274321313.00
## host_name*                         5431.95     4216.51    1.00     11453.00
## neighbourhood_group*                  2.61        1.48    1.00         5.00
## neighbourhood*                      106.81       88.96    1.00       221.00
## latitude                             40.73        0.05   40.50        40.91
## longitude                           -73.96        0.04  -74.24       -73.71
## room_type*                            1.48        0.00    1.00         3.00
## price                               121.43       68.20    0.00     10000.00
## minimum_nights                        3.58        2.97    1.00      1250.00
## number_of_reviews                    12.45        7.41    0.00       629.00
## last_review*                       1261.74      265.39    1.00      1765.00
## reviews_per_month                     1.06        0.92    0.01        58.50
## calculated_host_listings_count        1.50        0.00    1.00       327.00
## availability_365                     96.50       66.72    0.00       365.00
##                                       range  skew kurtosis        se
## id                              36484706.00 -0.09    -1.23  49669.87
## name*                              47905.00  0.00    -1.20     62.50
## host_id                        274318875.00  1.21     0.17 355509.26
## host_name*                         11452.00  0.05    -1.16     14.62
## neighbourhood_group*                   4.00  0.37    -0.11      0.00
## neighbourhood*                       220.00  0.26    -1.26      0.31
## latitude                               0.41  0.24     0.15      0.00
## longitude                              0.53  1.28     5.02      0.00
## room_type*                             2.00  0.42    -0.97      0.00
## price                              10000.00 19.12   585.59      1.09
## minimum_nights                      1249.00 21.83   853.95      0.09
## number_of_reviews                    629.00  3.69    19.53      0.20
## last_review*                        1764.00 -0.83    -1.02      3.17
## reviews_per_month                     58.49  3.13    42.49      0.01
## calculated_host_listings_count       326.00  7.93    67.54      0.15
## availability_365                     365.00  0.76    -1.00      0.60
```

Checking for missing values

```
summary(is.na(nycData))  ## checking Null values
```

```
##      id             name          host_id        host_name
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:48895     FALSE:48895     FALSE:48895     FALSE:48895
##
##  neighbourhood_group neighbourhood    latitude       longitude
##  Mode :logical       Mode :logical   Mode :logical   Mode :logical
##  FALSE:48895         FALSE:48895     FALSE:48895     FALSE:48895
##
##  room_type         price         minimum_nights  number_of_reviews
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:48895     FALSE:48895     FALSE:48895     FALSE:48895
##
##  last_review     reviews_per_month calculated_host_listings_count
##  Mode :logical   Mode :logical     Mode :logical
##  FALSE:48895     FALSE:38843       FALSE:48895
##                  TRUE :10052
##  availability_365
##  Mode :logical
##  FALSE:48895
##
```

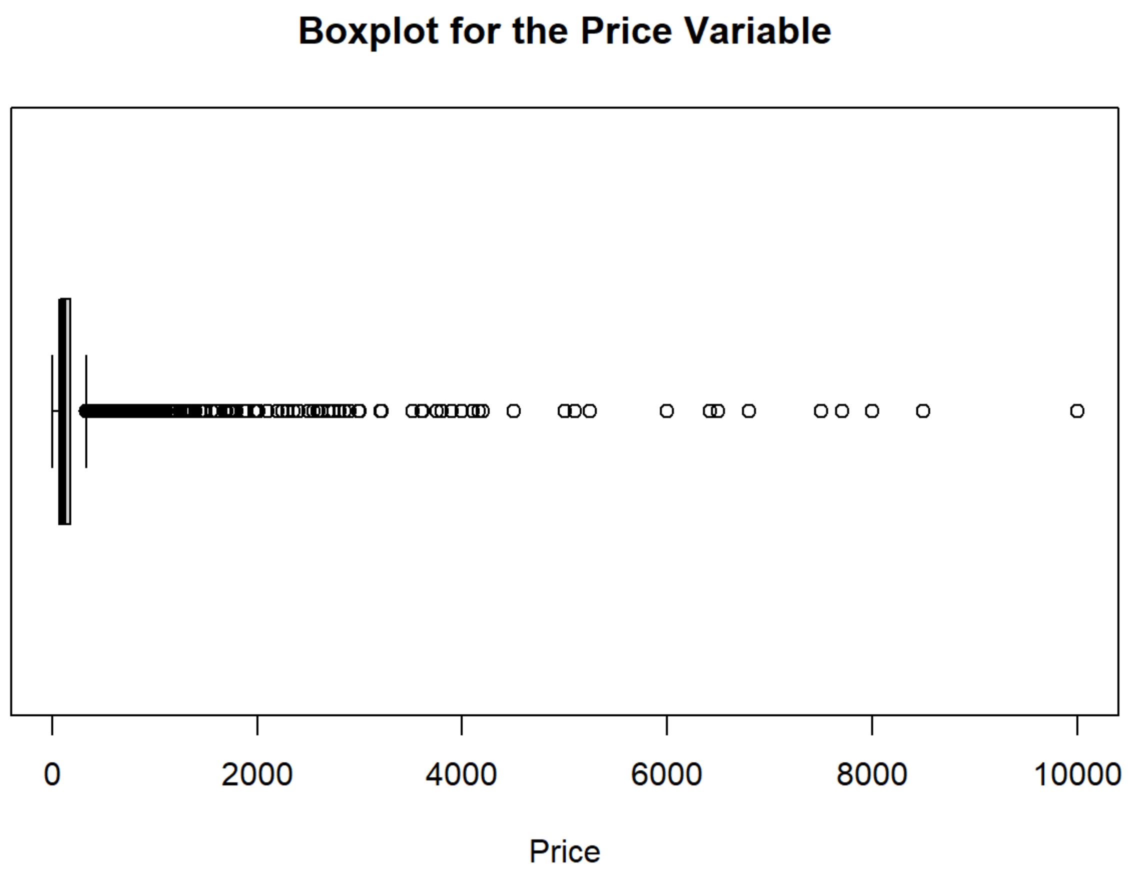
Here, The most important (target) variable is price. We can find out the some graphics of Price

```
library('ggplot2')
```

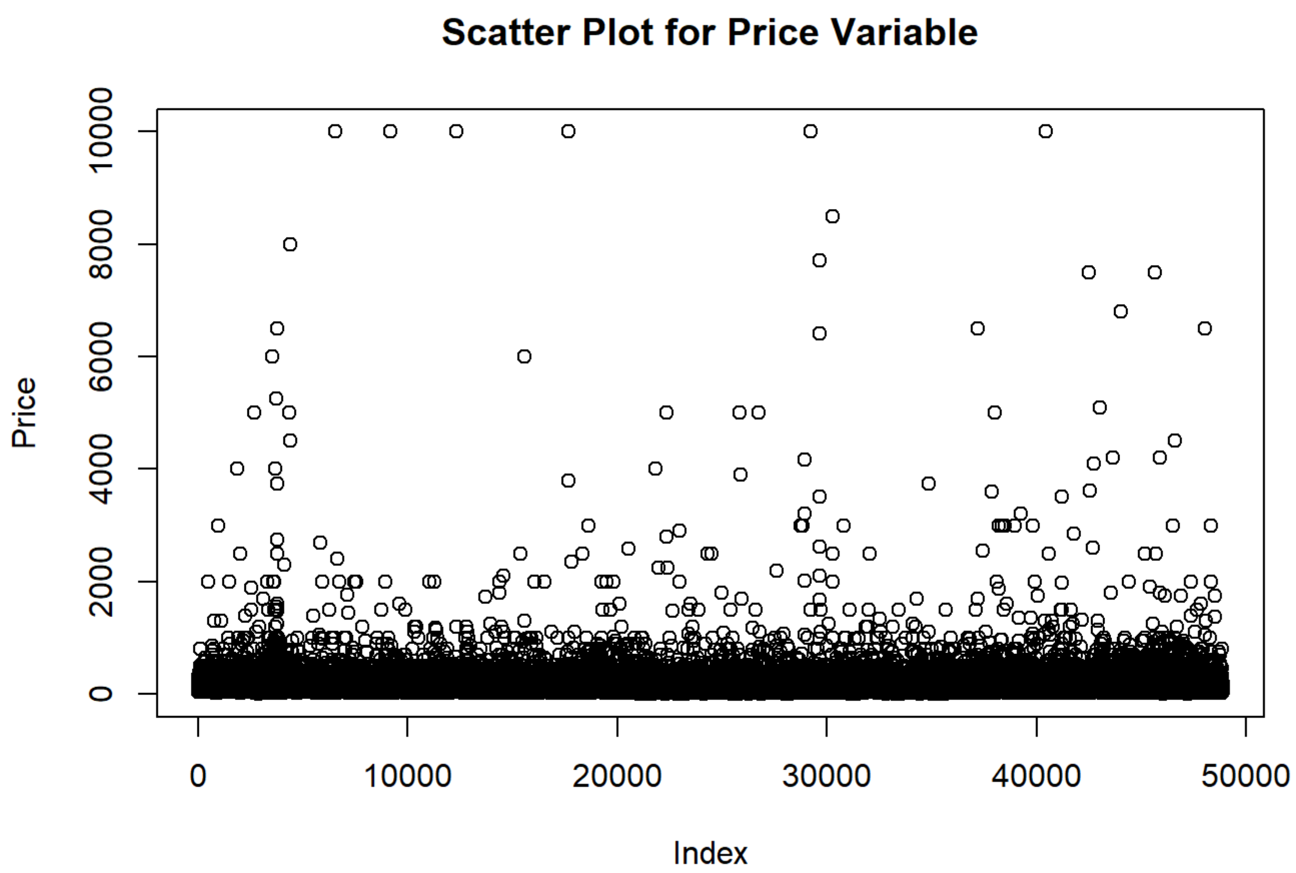
```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
boxplot(nycData$price, horizontal = TRUE, main="Boxplot for the Price Variable", xlab= "Pric
e")
```

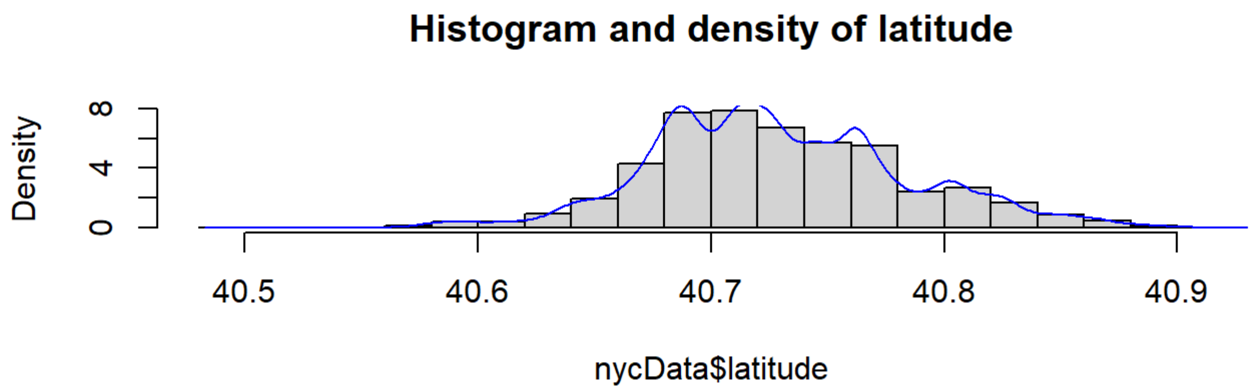


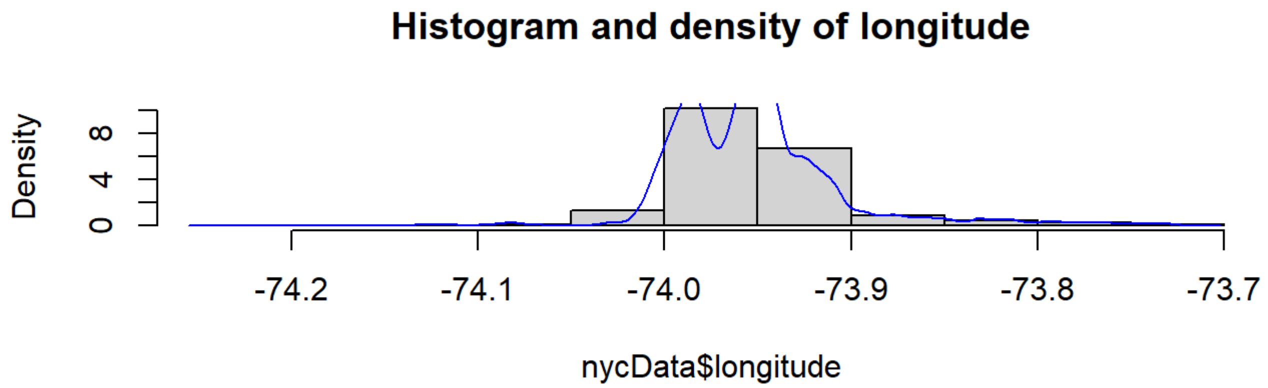

```
plot(nycData$price, main = "Scatter Plot for Price Variable", ylab= "Price")
```



we can also see the graphs for the latitude and longitude variable

```
par(mfrow=c(2,1))
hist(nycData$latitude, freq = FALSE, main= "Histogram and density of latitude")
lines(density(nycData$latitude), col="blue")

hist(nycData$longitude,  freq = FALSE, main= "Histogram and density of longitude")
lines(density(nycData$longitude), col="blue")
```






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

CHAPTER 2: Corelation of variables

In this chapter, we will try to find out the relationship between the variables. There are many numerical variables in our data set. We noticed that in our numerical data set, there are some missing values. If we give a look on *reviews_per_month* variable, there are missing values. Here, it is clear that *reviews_per_month* is missing only when number of reviews variable is 0 (Proof in below R code). So, we can replace the missing values by zero.

After handling the missing values, we can see the relationship among variables by plotting 2 graphs. From the graphs, we can understand the relationship of variables.

Question: Does price of the houses change linearly with any other factor and is there any linear relationship among the variables?

```
sum(nycData$number_of_reviews == 0)
```

```
## [1] 10052
```

```
# number_of_reviews is zero 10052 times, let's check whether number_of_reviews = 0 and review
s_per_month = NA both are happening together or not.
sum <- 0
for (i in 1:48895) {
  if (nycData$number_of_reviews[i] == 0 && is.na(nycData$reviews_per_month[i])) {
    sum <- sum + 1
  }
}
print(sum)
```

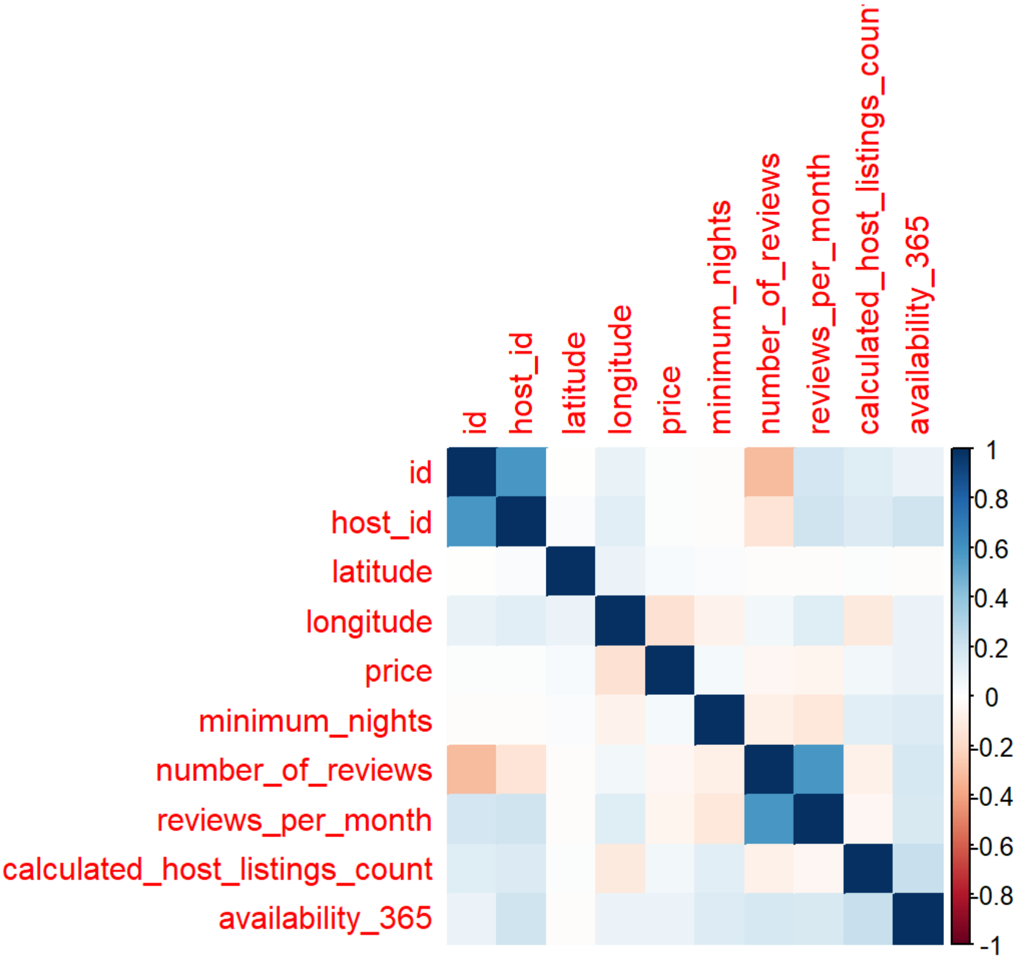
```
## [1] 10052
```

```
# So, both are happening together. Now let's replace.


nycData$reviews_per_month[is.na(nycData$reviews_per_month)] <- 0


# taking only the numerical variables of the nycData

nycNumerical <- subset(nycData, select = c(id, host_id, latitude, longitude, price, minimum_n
ights, number_of_reviews, reviews_per_month, calculated_host_listings_count, availability_36
5))

# correlation matrix
M <- cor(nycNumerical)

# plotting the correlation matrix
library(corrplot)
```

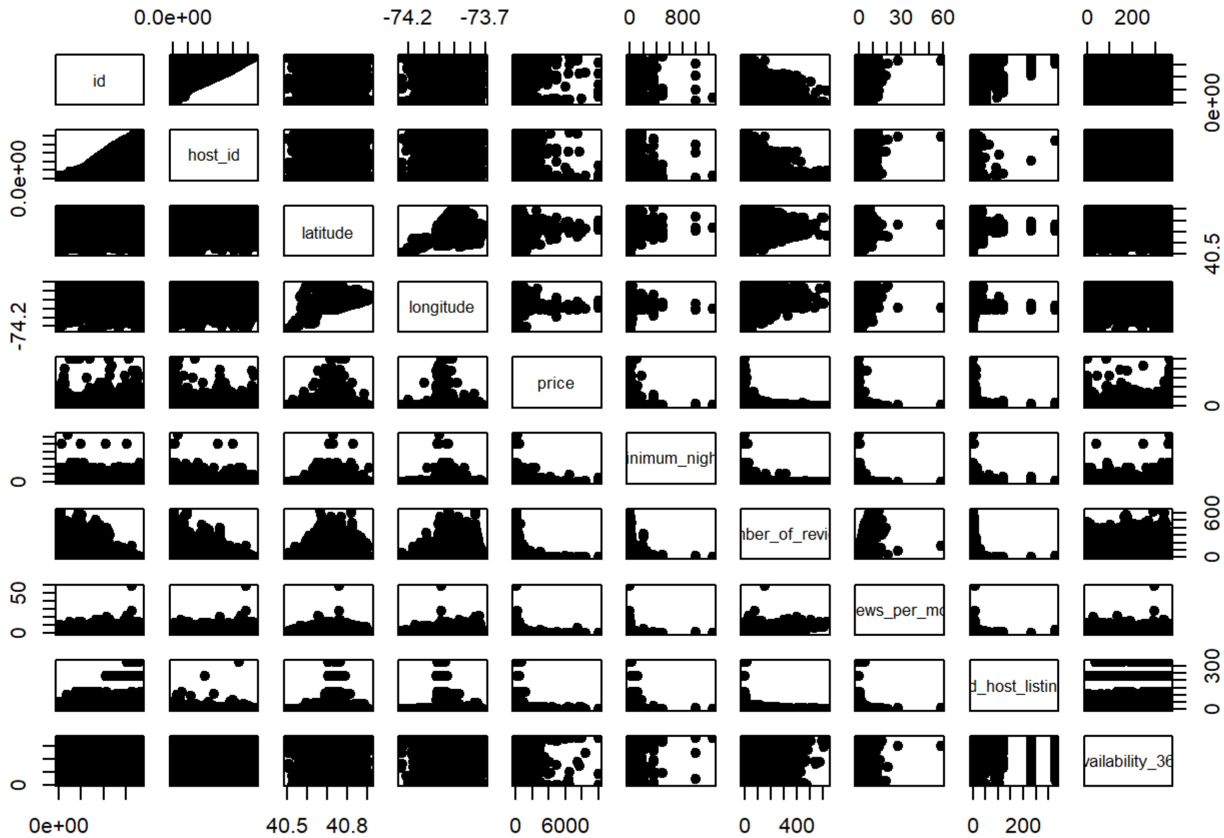
```
## Warning: package 'corrplot' was built under R version 4.0.5
```

```
## corrplot 0.89 loaded
```

```
corrplot(M, method = "color")
```




```
# our ques can be also answered by using scatterplot matrix

pairs(nycNumerical, pch = 19)
```



Above, we can see a correlation plot and a scatter plot matrix of the numerical variables of our data set.A Correlation Plot represents the strength of linear relation among the numerical variables.

We assumed at least some linear relation between the numerical variables initially(At least the price and the popularity of the housing). Here, the number_of_reviews variable is our way to understand the popularity of an Airbnb.

However, we don't see any significant relationship. So, we can say that the price of the housings is not linearly dependent on any other variable of our dataset and there is no significant linear relationship among the numerical variables of our data set.

CHAPTER 3: Location and Room Type

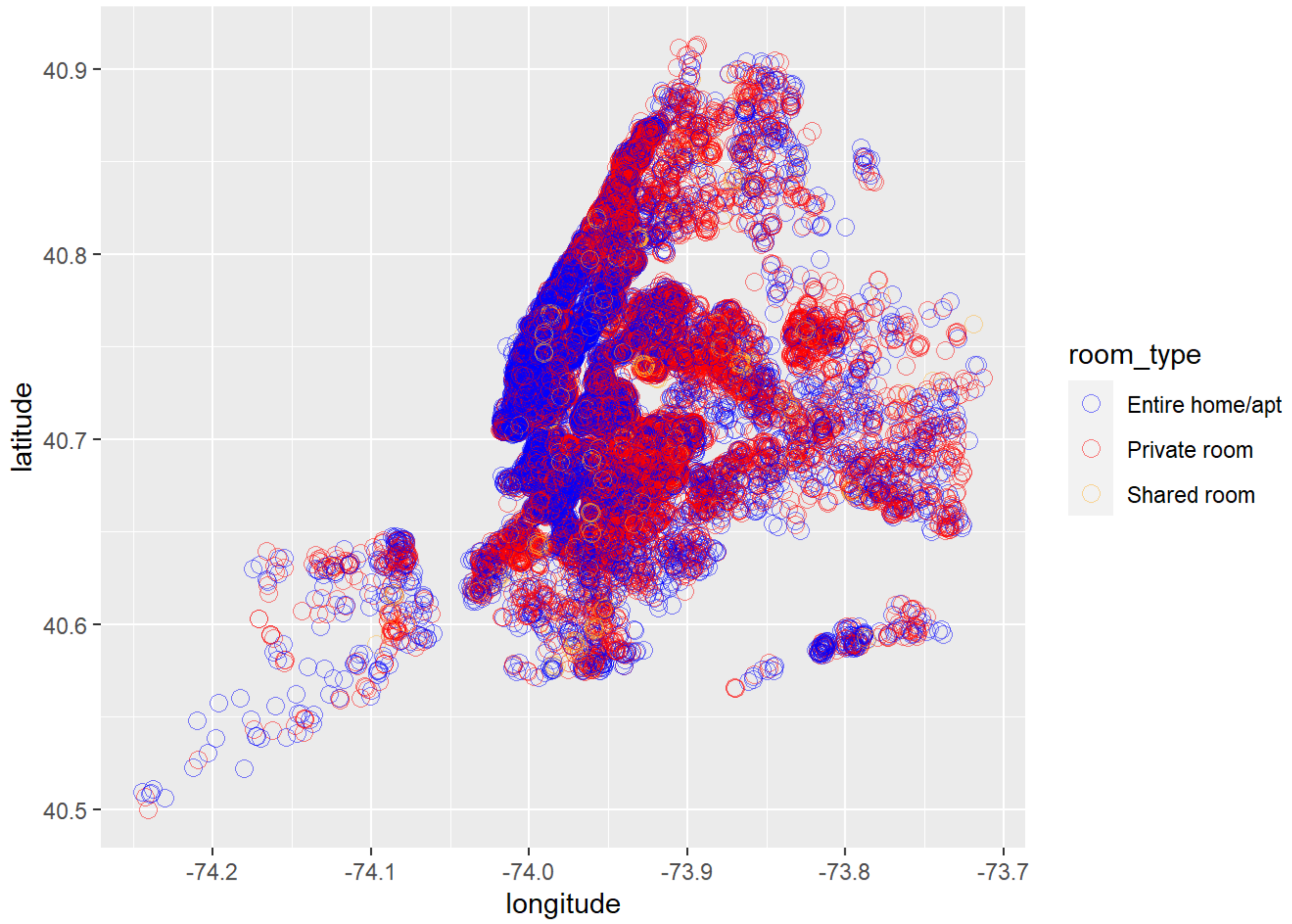
In our data set, we have 5 locations(cities) in New York. All location do not have same number of room type. In some areas the number of Entire Home is high and on the other hand some areas contain a lot of private rooms. There are some shared rooms also.

Does the geographic location have a significant effect on the type of room?

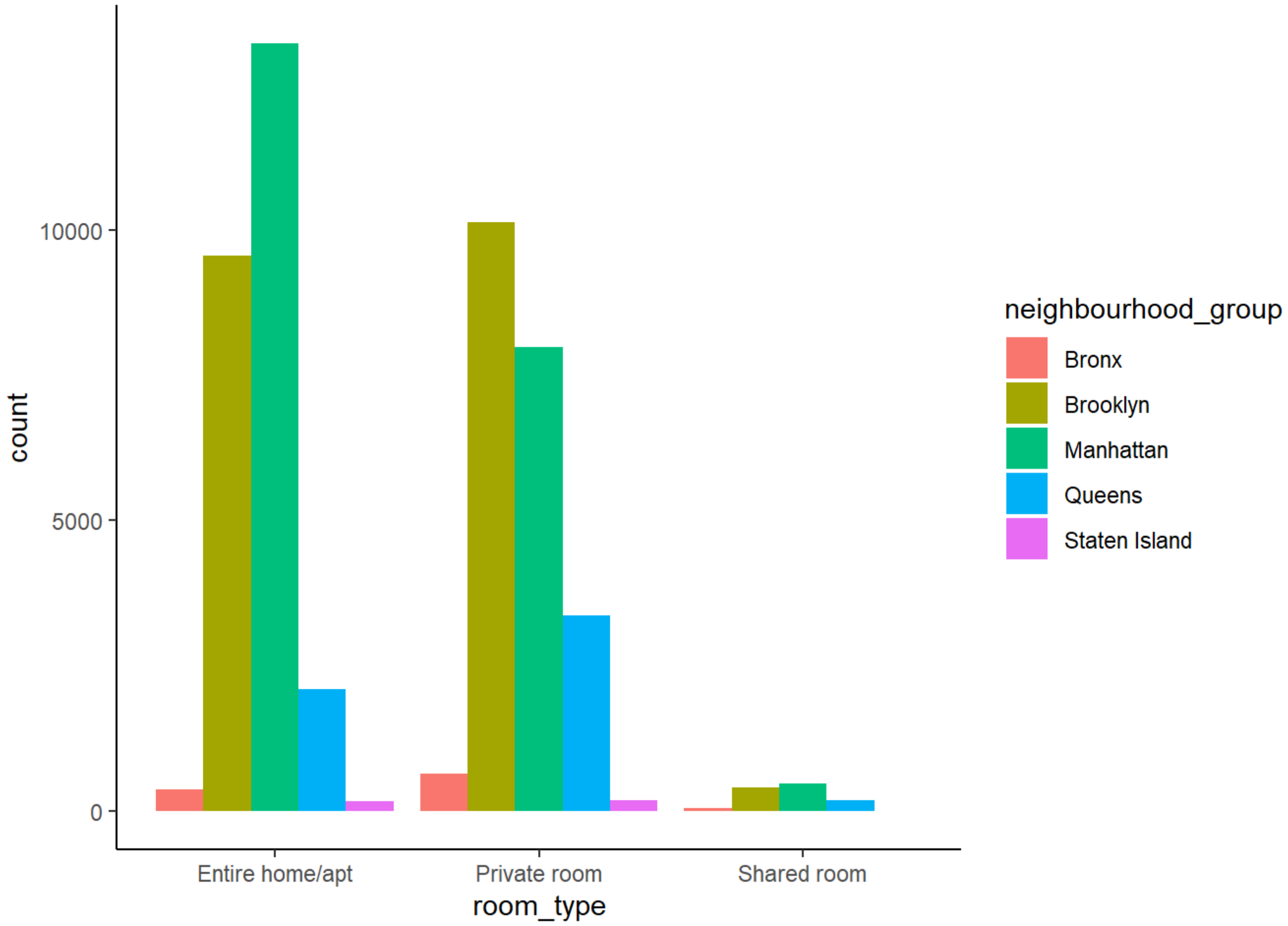
If we plot the location and room type, we see that in Manhattan, most of the room type is Entire Home.The percentage of Private rooms are almost same in all areas. But shared rooms are very few in every location.

We will also plot a bar diagram to understand better.

```
# Does the geographic location have a significant effect on the type of room?

library(ggplot2)

ggplot(nycData) +
  aes(x = longitude, y = latitude, color = room_type) +
  geom_point(shape = 21, alpha = 0.5, size = 3) +
  scale_color_manual(values = c("#0000FF","#FF0000", "#FFB533"))
```



```
# we can observe in a particular location there are more Entire home/apt (Longitude and Latit
ude range need to be mentioned)

#bar plot can also answer our question

ggplot(nycData, aes(x = room_type, fill = neighbourhood_group)) +
  geom_bar(position = position_dodge()) +
  theme_classic()
```



Above, there are two plots. One is a scatter plot and the other one is a bar diagram. We can see from the scatter plot in a particular range of longitude and latitude (approximately latitude 40.5 to 40.8 and longitude -74.05 to -73.95), Entire home is more dense. However, we can't be completely sure because of the overlapping between the points. So, we will try to visualize by using a bar plot.

Next, if we consider the above bar diagram we will get a very clear about the in room type in different cities. Most of the entire homes and private room are situated in Manhattan and Brooklyn according to our bar plot.

CHAPTER 4: Location And Price

Price is a very important variable in our data set. There might be some places where the price is more than other places and we are very interested in finding that. We will try to find the expensive places and the cheap places depending on the price of the housings.

We can make a question here

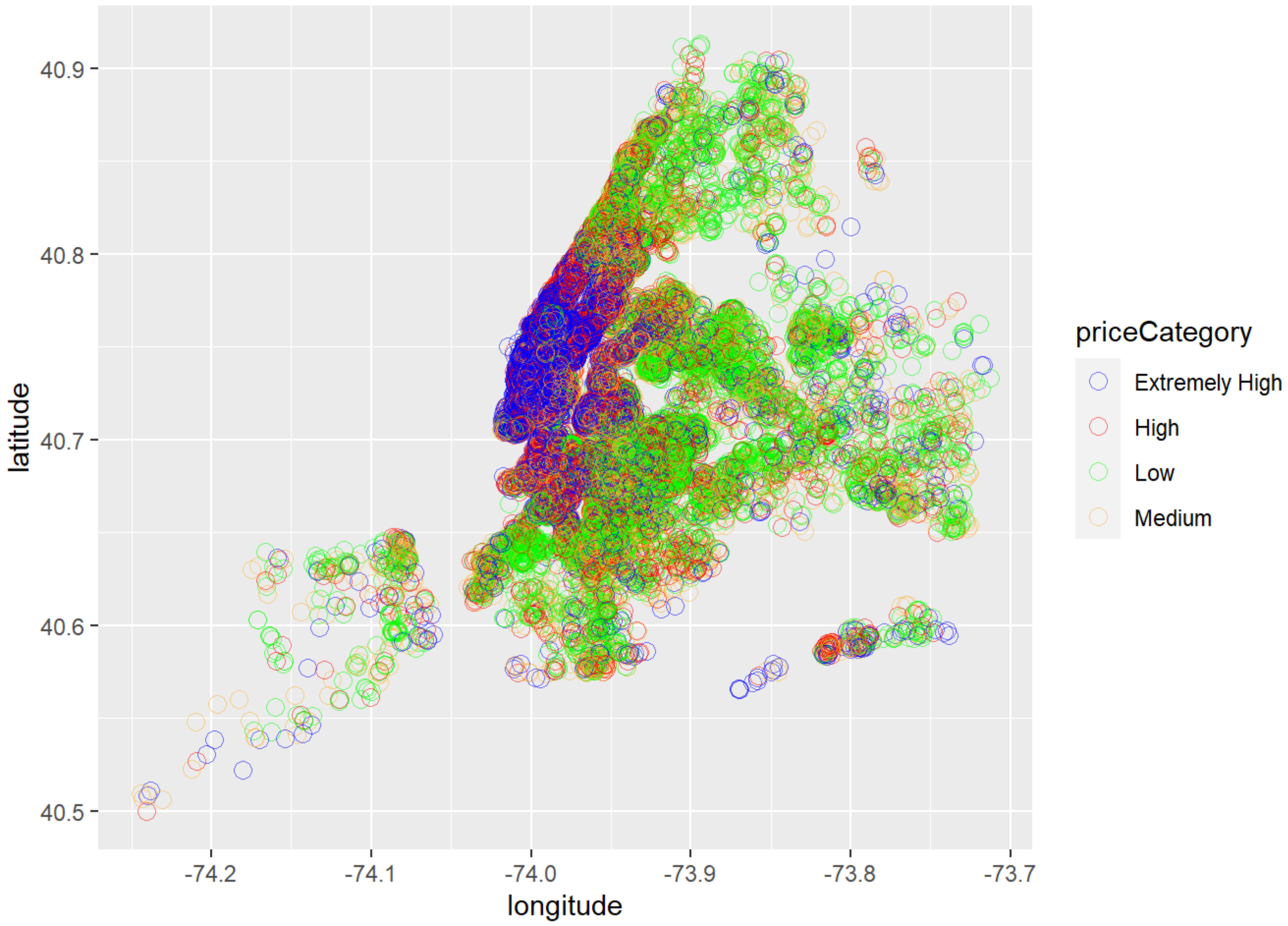
Question: Does location has an impact on price?

Yes, location has an impact on price. We have categorized the the price variables in 4 Categories(Extremely High, High, Low and Medium). We used quantile method to find out the ranges. Then, we plotted two graphics for our analysis.

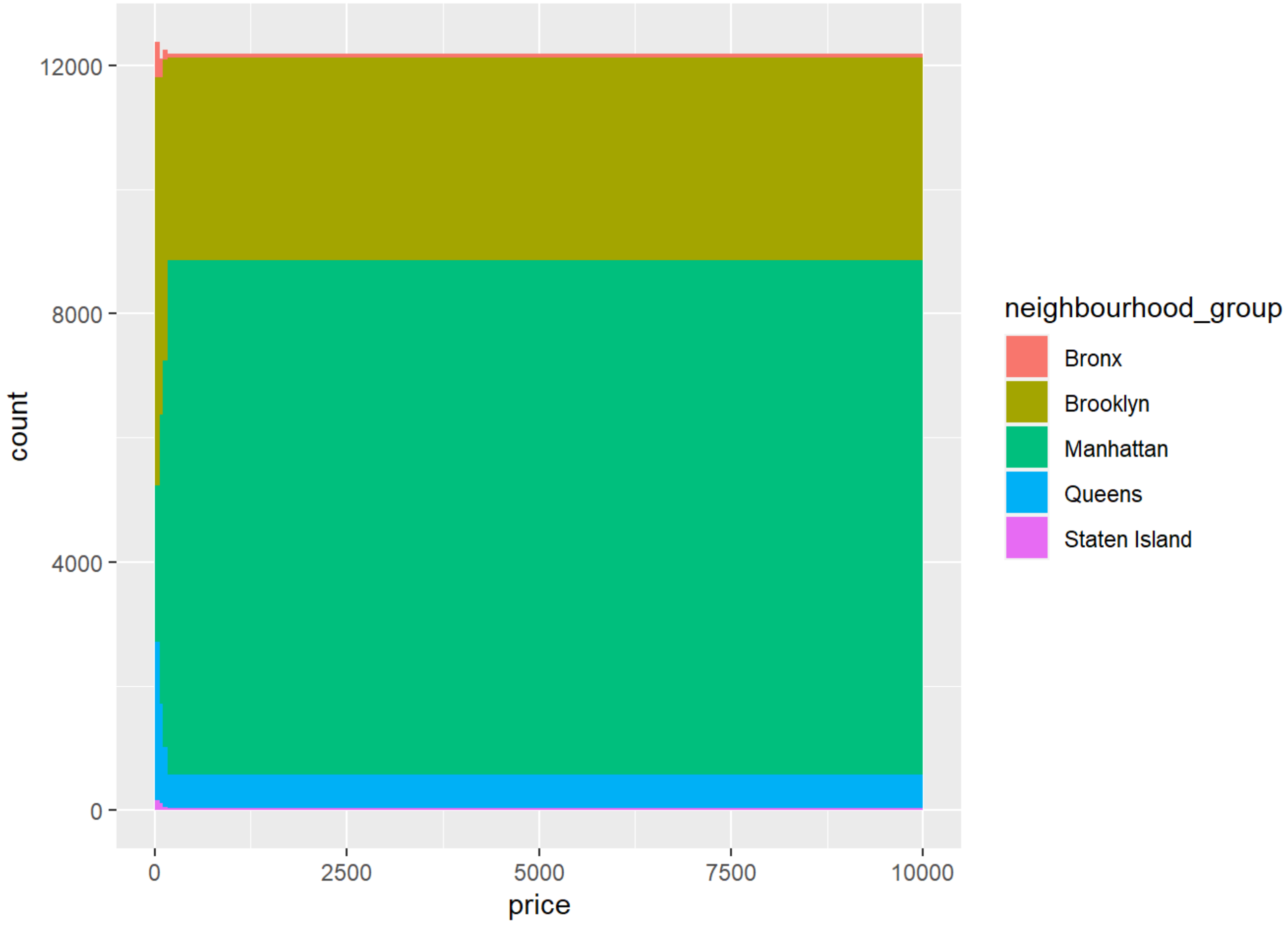
```
#ques:  Does location has an impact on price?

quantile(nycData$price)
```

```
##    0%   25%   50%   75%  100% 
##     0    69   106   175 10000
```

```
# Categorize the price variable
for (i in 1:48895) {
  if (nycData$price[i] <= 69) {
    nycData$priceCategory[i] <- "Low"
  } else if (69 < nycData$price[i] & nycData$price[i] <= 106) {
    nycData$priceCategory[i] <- "Medium"
  } else if (nycData$price[i] > 106 & nycData$price[i] <= 175) {
    nycData$priceCategory[i] <- "High"
  } else {
    nycData$priceCategory[i] <- "Extremely High"
  }
}



library(ggplot2)

#scatterplot
ggplot(nycData) +
  aes(x = longitude, y = latitude, color = priceCategory) +
  geom_point(shape = 21, alpha = 0.5, size = 3) +
  scale_color_manual(values = c("#0000FF","#FF0000", "#00FF00", "#FFB533"))
```




```
#histogram
ggplot(nycData, aes(x = price, fill = neighbourhood_group)) +
  geom_histogram(boundary = 0, breaks = c(0 ,69, 106, 175, 10000))
```



Upwards, we see two plots. In the scatter plot, we can see that in a particular range(approximately latitude 40.5 to 40.8 and longitude -74.0 to -73.9) of longitude and latitude there are more high price housings. But, again we can't say that with complete confidence because of the overlapping. We will see a histogram for our conclusion.

So, if we look at the histogram we can see that Manhattan and Brooklyn is the most expensive places than any other cities.

Histogram & Density with log10 transformation for neighbourhood areas

New York City consist of five neighbourhood areas:

1. Manhattan
2. Brooklyn
3. Queens
4. The Bronx
5. Staten Island.

It can be useful to vizualise the distribution of price for every neighbourhood area.

```
library(magrittr)
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
```

```
airbnb_nh <- nycData %>%
  group_by(neighbourhood_group) %>%
  summarise(price = round(mean(price), 2))
```

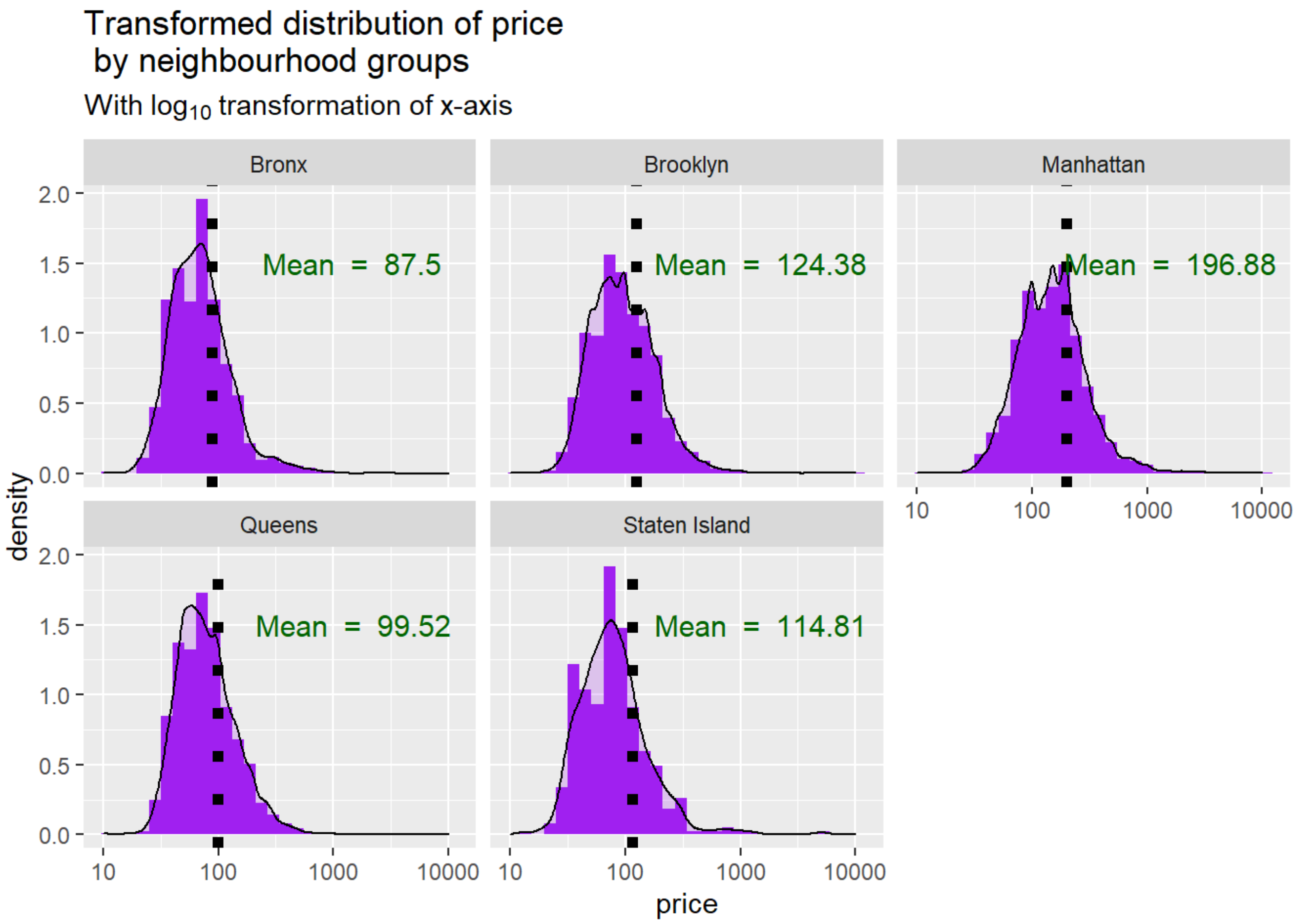
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
ggplot(nycData, aes(price)) +
  geom_histogram(bins = 30, aes(y = ..density..), fill = "purple") +
  geom_density(alpha = 0.2, fill = "purple") +
  ggtitle("Transformed distribution of price\n by neighbourhood groups",
          subtitle = expression("With" ~'log'[10] ~ "transformation of x-axis")) +
  geom_vline(data = airbnb_nh, aes(xintercept = price), size = 2, linetype = 3) +
  geom_text(data = airbnb_nh,y = 1.5, aes(x = price + 1400, label = paste("Mean  = ",price)),
color = "darkgreen", size = 4) +
  facet_wrap(~neighbourhood_group) +
  scale_x_log10()
```

```
## Warning: Transformation introduced infinite values in continuous x-axis

## Warning: Transformation introduced infinite values in continuous x-axis
```

```
## Warning: Removed 11 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 11 rows containing non-finite values (stat_density).
```



we have 5 neighborhood areas in our data set. For each area the prices are not same that is why we tried to find out the average price for each area. It is clear from the graph that Manhattan is the most expensive area and Brooklyn is the second highest expensive. It is visible from the graph that The Bronx is cheapest area.

CHAPTER 5: Popularity of Room

In this section, we will try to illustrate popularity based on the price of the housings.

If we make question like this for our dataset:

Question: Which type of room is the cheapest and is that more popular among the customers?

We will now try to find the answer from some visualizations.

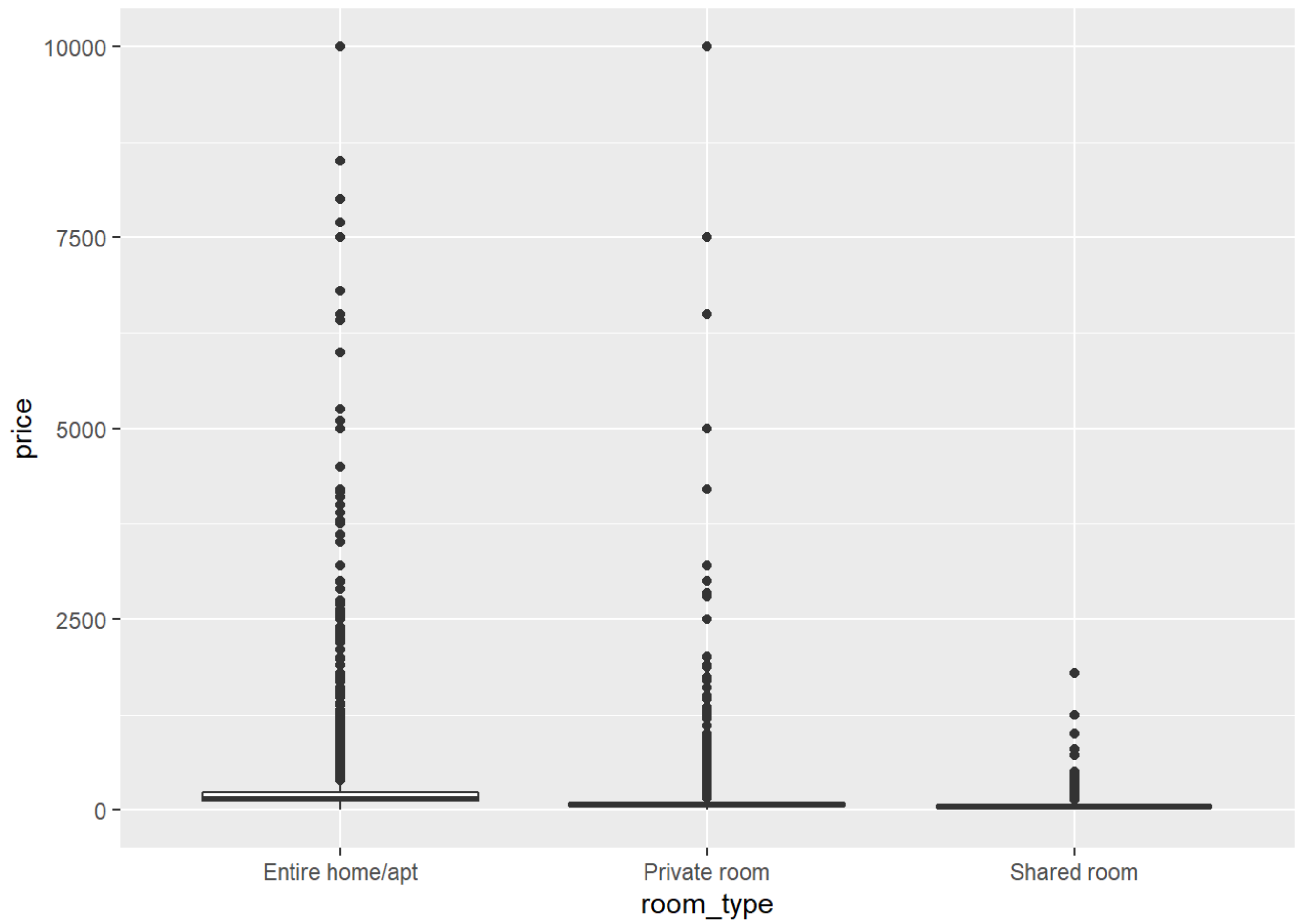
```
#Which type of room is the cheapest and is that more popular among the customers?
library(ggplot2)
library(gplots)
```

```
## Warning: package 'gplots' was built under R version 4.0.5
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
#check which one is the cheapest
ggplot(nycData, aes(x=room_type, y=price)) +
  geom_boxplot()
```




```
#mean plot to be extra sure
plotmeans(price ~ room_type, data = nycData, frame = FALSE,
          mean.labels = TRUE, connect = FALSE)
```

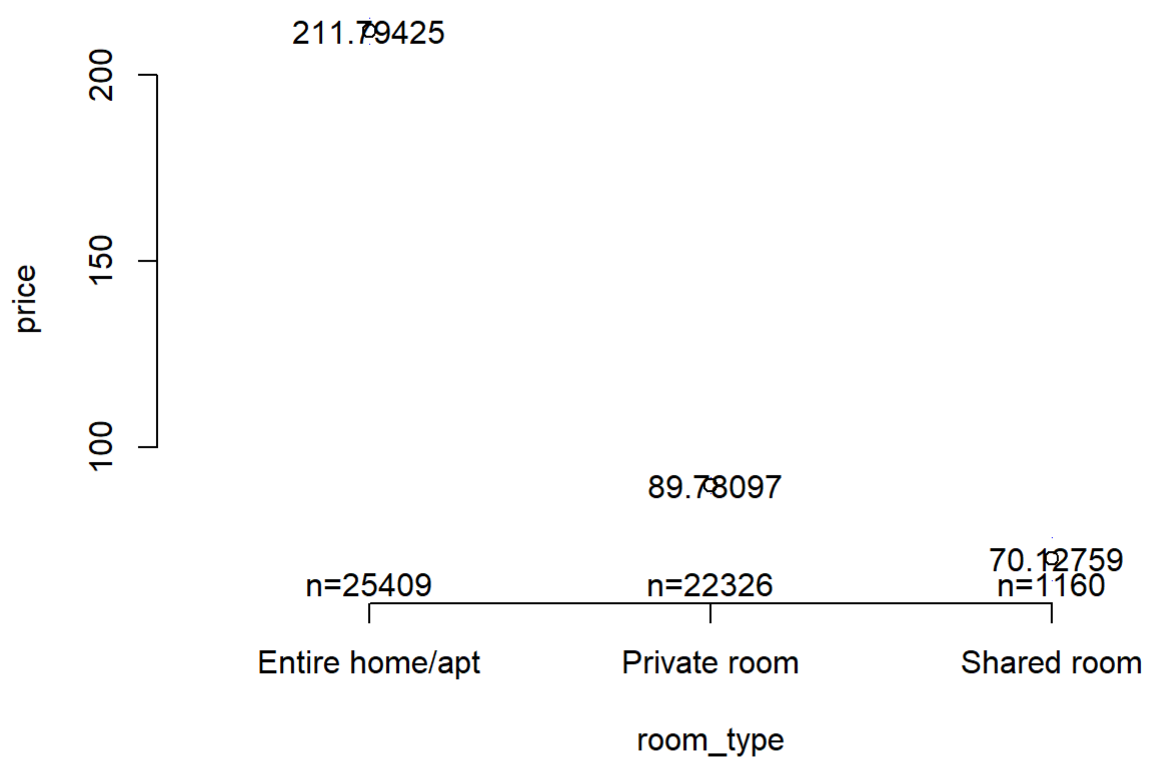
```
## Warning in text.default(x, y, label = labels, col = col, ...): "frame" is not a
## graphical parameter
```

```
## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped
```

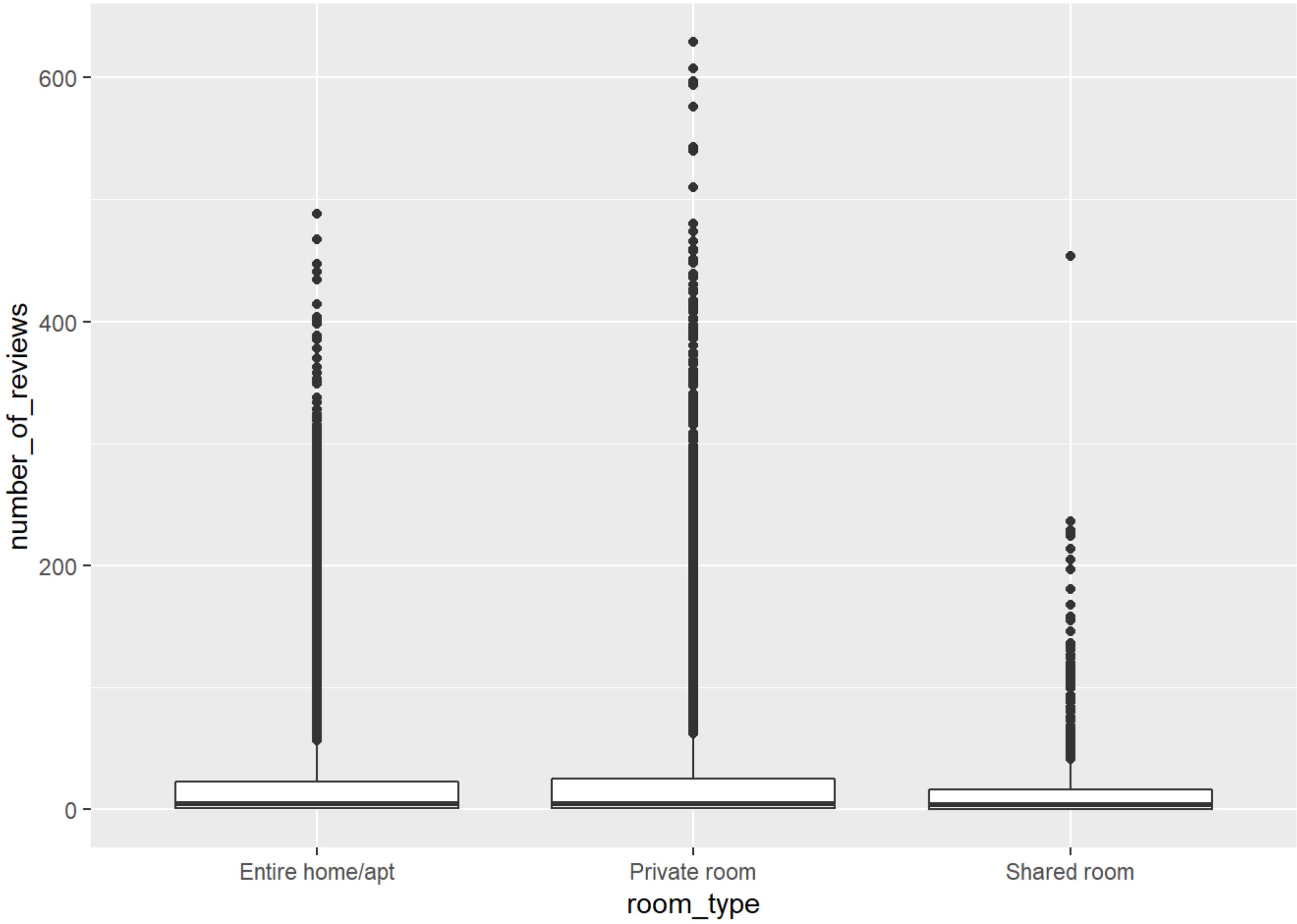
```
## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" is not a
## graphical parameter
```

```
## Warning in axis(1, at = 1:length(means), labels = legends, ...): "frame" is not
## a graphical parameter
```




```
#We found that shared room is the cheapest
#Now let's see is that most popular among the customers of arbnb

ggplot(nycData, aes(x=room_type, y=number_of_reviews)) +
  geom_boxplot()
```

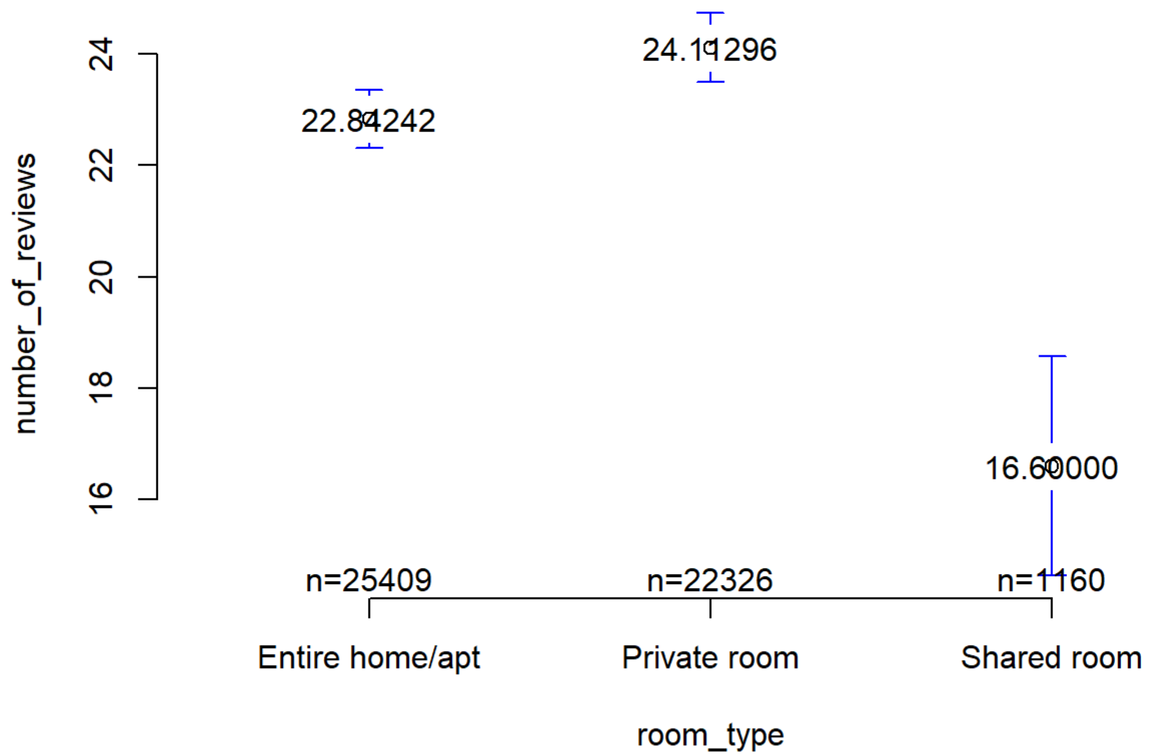



```
#mean plot to be extra sure
plotmeans(number_of_reviews ~ room_type, data = nycData, frame = FALSE,
          mean.labels = TRUE, connect = FALSE)
```

```
## Warning in text.default(x, y, label = labels, col = col, ...): "frame" is not a
## graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" is not a
## graphical parameter
```

```
## Warning in axis(1, at = 1:length(means), labels = legends, ...): "frame" is not
## a graphical parameter
```



Upwards, we can see 4 graphics. First one is a box plot and from this box plot we can see that shared room is most probably the cheapest one. However to be totally sure about the fact we will have a look at the second graph. The second graph is representing the mean price of each type of housing. Here, we also see that the mean of shared room is less than the other two. Hence, we can say that the shared room is the cheapest.

Now another important question comes in our mind! Is shared room is the most popular one among the customers?

Our remaining plots provide the answer of the question. We assumed that the number of reviews represents the popularity.

From our third plot we see that, the shared room is most likely not the most popular one. Again, to have a concrete idea we have to go to the mean plot. From the 4th plot, it can seen that the mean number of reviews for the shared room is the minimum. So, it is definitely not the most popular one.

From the above discussion and plots, we would like to say that the cheapest one is not the most popular one here in our dataset.

Chapter 6: Final Analysis

In this chapter, we will try to understand a very interesting thing. At this point of our discussion, we can easily say that Manhattan is most likely the most expensive area.

At this moment, we would like to check if the rent is decreasing as we go far from Manhattan.

We will calculate the distance between 2 points by using the Haversine Formula. So, we would like to present the Haversine Formual first-

This uses the 'haversine' formula to calculate the great-circle distance between two points – that is, the shortest distance over the earth's surface – giving an 'as-the-crow-flies' distance between the points (ignoring any hills they fly over, of course!).

Haversine formula: $a = \sin^2(\Delta\varphi/2) + \cos \varphi 1 \cdot \cos \varphi 2 \cdot \sin^2(\Delta\lambda/2)$ $c = 2 \cdot \operatorname{atan2}(\sqrt{a}, \sqrt{(1-a)})$ $d = R \cdot c$ where φ is latitude, λ is longitude, R is earth's radius (mean radius = 6,371km); note that angles need to be in radians to pass to trig functions!

```
nycData$longitudeRadian <- nycData$longitude * (3.1416/180)

nycData$latitudeRadian <- nycData$latitude * (3.1416/180)


#Manhattan's longitude and latitude

# 40.78343, -73.96625

manhattanLongitude <- -73.96625*(3.1416/180)

manhattanLatitude <- 40.78343*(3.1416/180)


#Haversine formula

for(i in 1:48895){
   a <- (sin((manhattanLatitude - nycData$latitudeRadian[i])/2))^2 + (cos(manhattanLatitude)*
 cos(nycData$latitudeRadian[i])*(sin((manhattanLongitude - nycData$longitudeRadian[i])/2))^2)
  c <- 2 * atan2(sqrt(a), sqrt(1-a))
  nycData$distanceFromManhattan[i] <- 6371*c


}


# Categorize the distanceFromManhattan variable
for (i in 1:48895) {
  if (nycData$distanceFromManhattan[i] <= 4.4676940    ) {
    nycData$distanceFromManhattanCat[i] <- "0 to 4.46 km"
   } else if (nycData$distanceFromManhattan[i] > 4.4676940 & nycData$distanceFromManhattan[i]
 <= 7.6721030 ) {
    nycData$distanceFromManhattanCat[i] <- "4.46 to 7.67 km"
  } else if (nycData$distanceFromManhattan[i] > 7.6721030  & nycData$distanceFromManhattan[i]
<= 11.1516474 ) {
    nycData$distanceFromManhattanCat[i] <- "7.67 to 11.15 km"
  } else {
    nycData$distanceFromManhattanCat[i] <- "More than 11.15 km"
  }
}


library(gplots)

plotmeans(price ~ distanceFromManhattanCat, data = nycData, frame = FALSE,
          mean.labels = TRUE, connect = TRUE)
```

```
## Warning in text.default(x, y, label = labels, col = col, ...): "frame" is not a
## graphical parameter
```

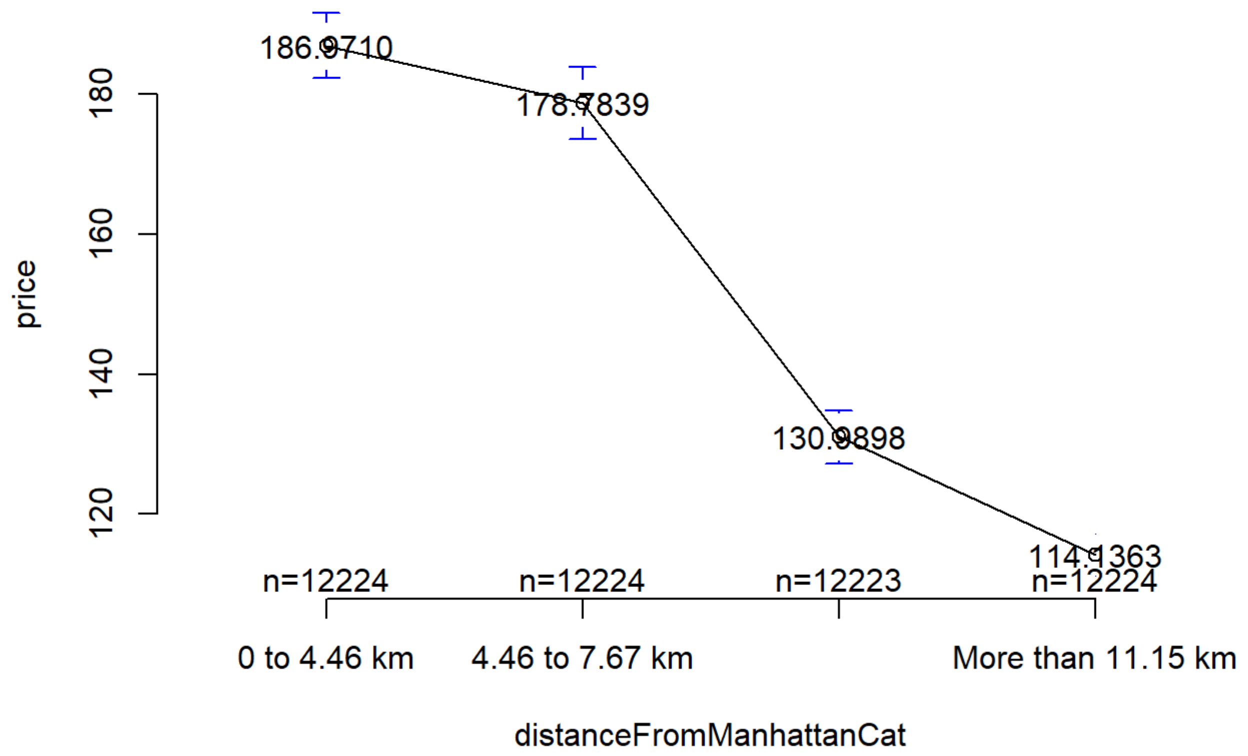
```
## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped
```

```
## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" is not a
## graphical parameter
```

```
## Warning in axis(1, at = 1:length(means), labels = legends, ...): "frame" is not
## a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" is not a
## graphical parameter
```



In the above code, first we calcaculated the distance of each Airbnb from Manhattan. Then we categorized each distance using quantiles. Then, we plotted the mean for each range.

From the Mean Plot we can easily say that the rent of the houses decreases as we go far from Manhattan.

CHAPTER 7 : Summary

Through this exploratory visualization project, we gained several interesting insights of the Airbnb rental market. Below we will summarize the answers of the questions that we wished to answer at the beginning of the project:

Question: Does price of the houses change linearly with any other factor and is there any linear relationship among the variables?

=> Price doesn’t change linearly with any other variables of our dataset and there is no linear relationship among the variables of our dataset.

Question: Does the geographic location have a significant effect on the type of room?

=> Yes. Most of the entire homes and private room are situated in Manhattan and Brooklyn.

Question: Does location has an impact on price?

=> Yes. Manhattan and Brooklyn is the most expensive places than any other cities.

Question: Which type of room is the cheapest and is that more popular among the customers?

=> Shared room is the cheapest and that is not the most popular one among the customers.

Question: Is the rent of the housing decreases as we go far from Manhattan?

=> Yes, the rent decreases as we go far from Manhattan.

References:

[1].https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

[2].https://www.kaggle.com/brittabettendorf/berlin-airbnb-data

[3].https://rpubs.com/Dkanawat/652521

[4].https://ggplot2.tidyverse.org/reference/

[5].https://www.airbnb.com/new-york-ny/stays