EDA Project 1

Tyler

2022-10-05

```
library(tidyverse)
library(seriation)
library(tidyquant)
library(tinytex)
```

```
airbnb <- read.csv("~/Airbnb_Open_data.csv")
```

```
airbnb <- airbnb %>%
  select(id, host.id,neighbourhood.group,neighbourhood,room.type,Construction.year,price,
         service.fee,minimum.nights, number.of.reviews, review.rate.number, availability.365)
```

Chapter 1: Data Descriptions

The following data is data taken from Airbnb in the New York City. This data is from the year 2019-2022. The variables in the data consist of ID, Name, host.id, host_identify_verified, host.name, neighbourhood.group, neighbourhood, lat, long, country, country.code, instant_bookable, cancellation_policy, room.type, construction.year, price, service.fee, minium.nights,number.of.reviews, last.review, reviews.per.month, review.rate.number, caculated.host.listings.count, availability.365, and house_rules. The price and service.fee are in USD.

Chapter 2: EDA

-

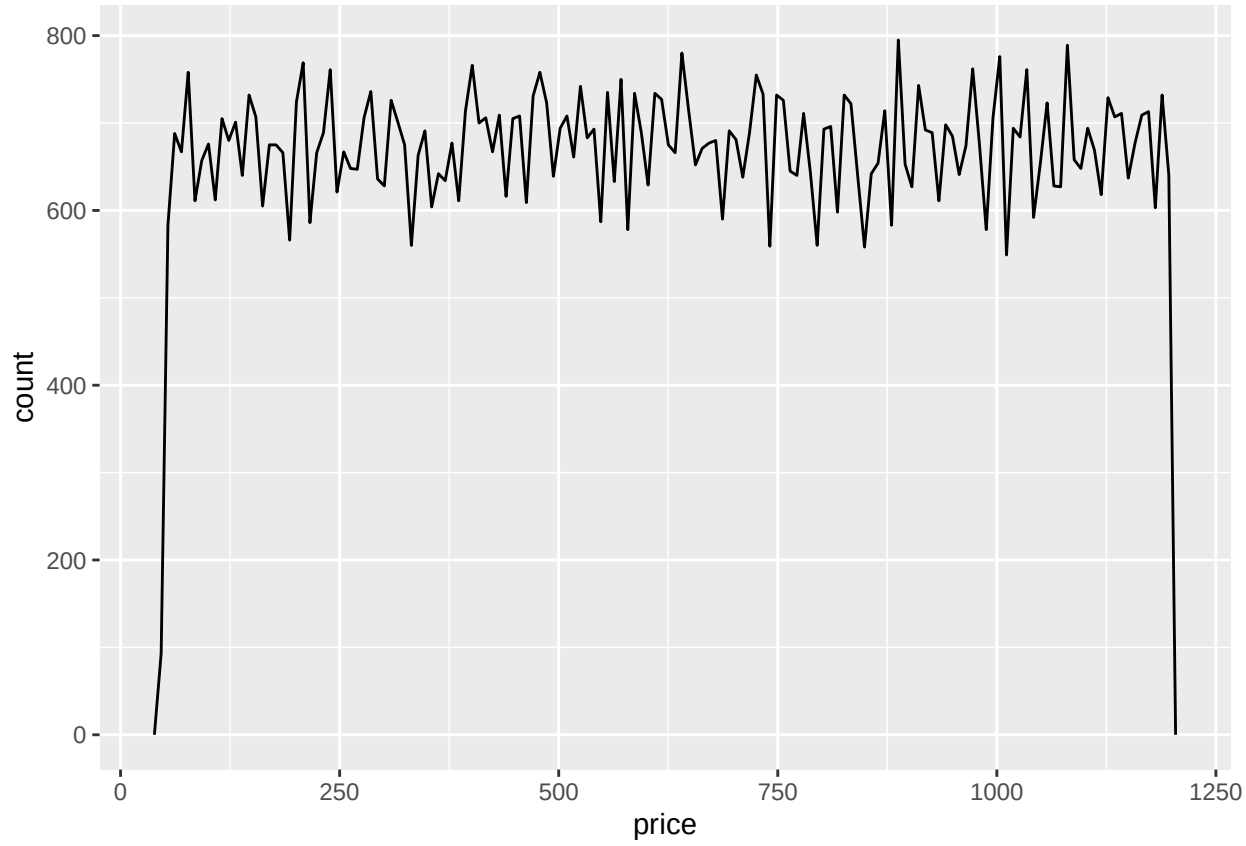
Section 1: I will be finding the center, spread, shape of the distribution for each variable. I also will be Checking for outliers

Price

```
airbnb %>%
  pull(price) %>%
  psych::describe()
```

```
##    vars      n   mean     sd median trimmed    mad min  max range skew kurtosis
## X1    1 102352 625.29 331.67    624  625.33 425.51  50 1200  1150    0    -1.19
##      se
## X1 1.04
```

```
airbnb %>% drop_na() %>%
  ggplot(mapping = aes(x = price)) +
  geom_freqpoly(bins = 150)
```




```
p50 <- airbnb %>%
  filter(price == 50)
```

I noticed that there was abnormal low value for price. To make sure it was not an outlier I wanted to check the room type. The room type listed for it is a private room which would make since for why it is only $50 a night. Some have them listed as an entire home/apartment. This could have something to do with the location of the Airbnb listed. This would require further investigation. Also the distribution is pretty symmetrical.

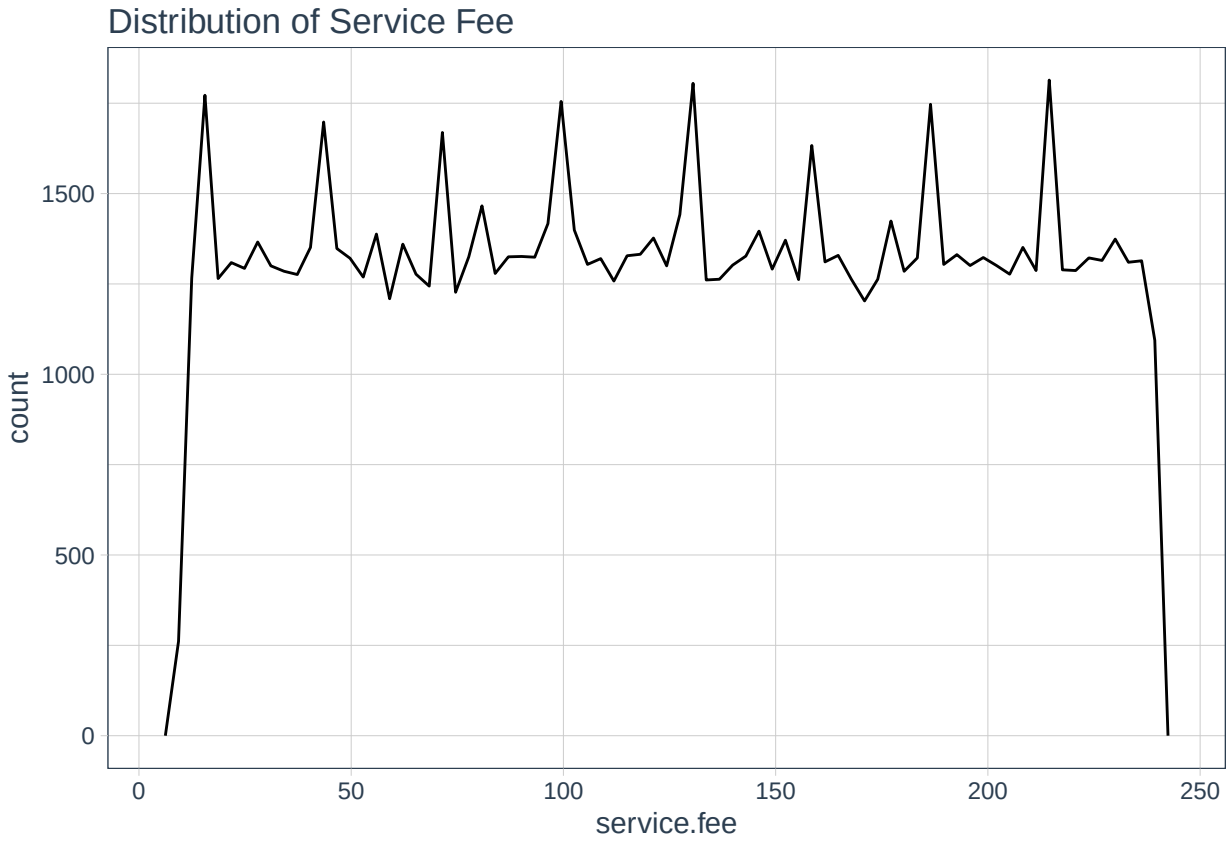
Service Fee

```
airbnb %>%
  pull(service.fee) %>%
  psych::describe()
```

```
##    vars      n   mean    sd median trimmed   mad min max range skew kurtosis
## X1    1 102326 125.03 66.33    125  125.03 84.51  10 240   230    0    -1.19
##      se
## X1 0.21
```

```
airbnb %>% drop_na() %>%
  ggplot(mapping = aes(x = service.fee)) +
  geom_freqpoly(bins = 75)+
  labs(title = "Distribution of Service Fee") +

  theme_tq()
```



Seems that distribution is pretty symterical for service fee.

Number of Reviews

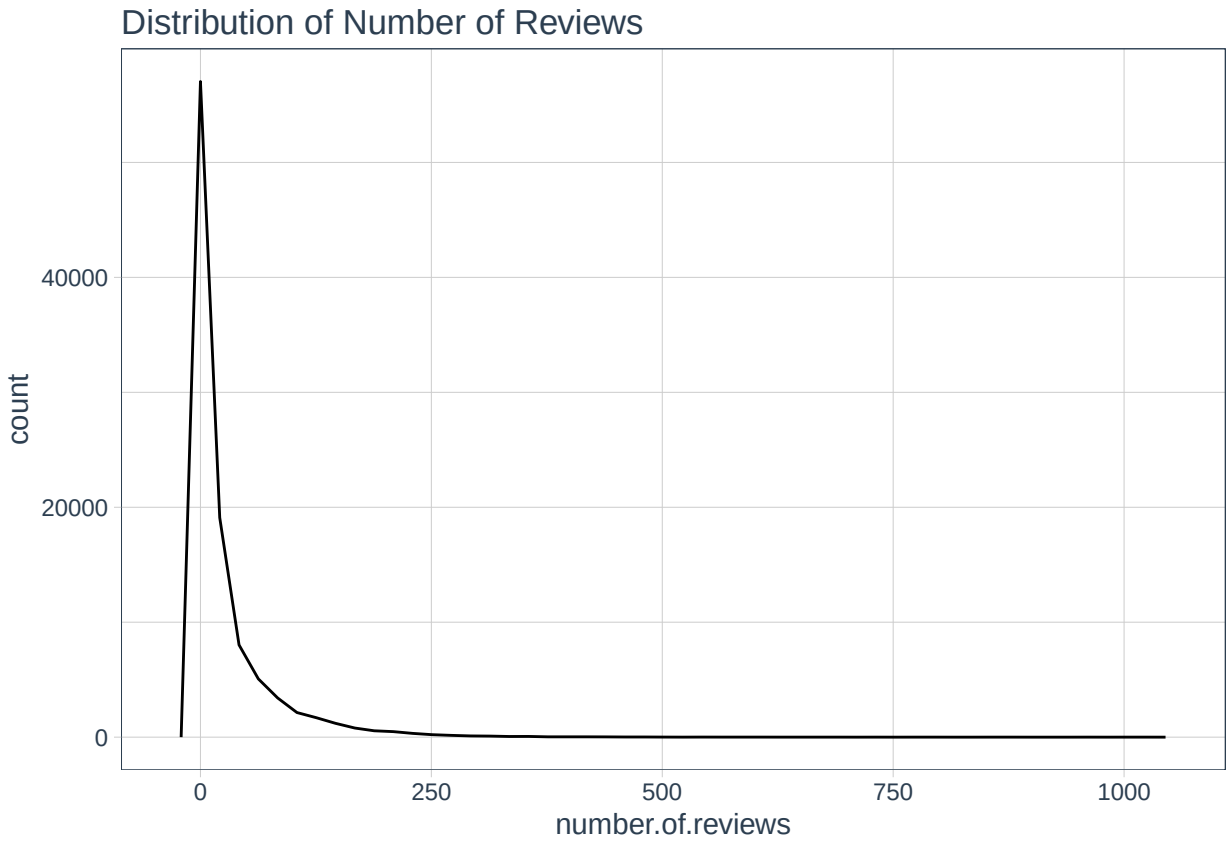
```
airbnb %>%
  pull(number.of.reviews) %>%
  psych::describe()
```

```
##    vars      n  mean    sd median trimmed   mad min  max range skew kurtosis
## X1    1 102416 27.48 49.51      7   15.93 10.38   0 1024  1024 3.84    25.03
##      se
## X1 0.15
```

```
airbnb %>% drop_na() %>%
  ggplot(mapping = aes(x = number.of.reviews)) +
  geom_freqpoly(bins = 50) +

labs(title = "Distribution of Number of Reviews") +

  theme_tq()
```



The distribution skews very left. This is probably do to the fact that many people leave 0 reviews.

Neighourhood Group

```
airbnb %>%
  count(neighbourhood.group) %>%
  mutate(percentage = (n/sum(n)) %>% scales::percent())
```

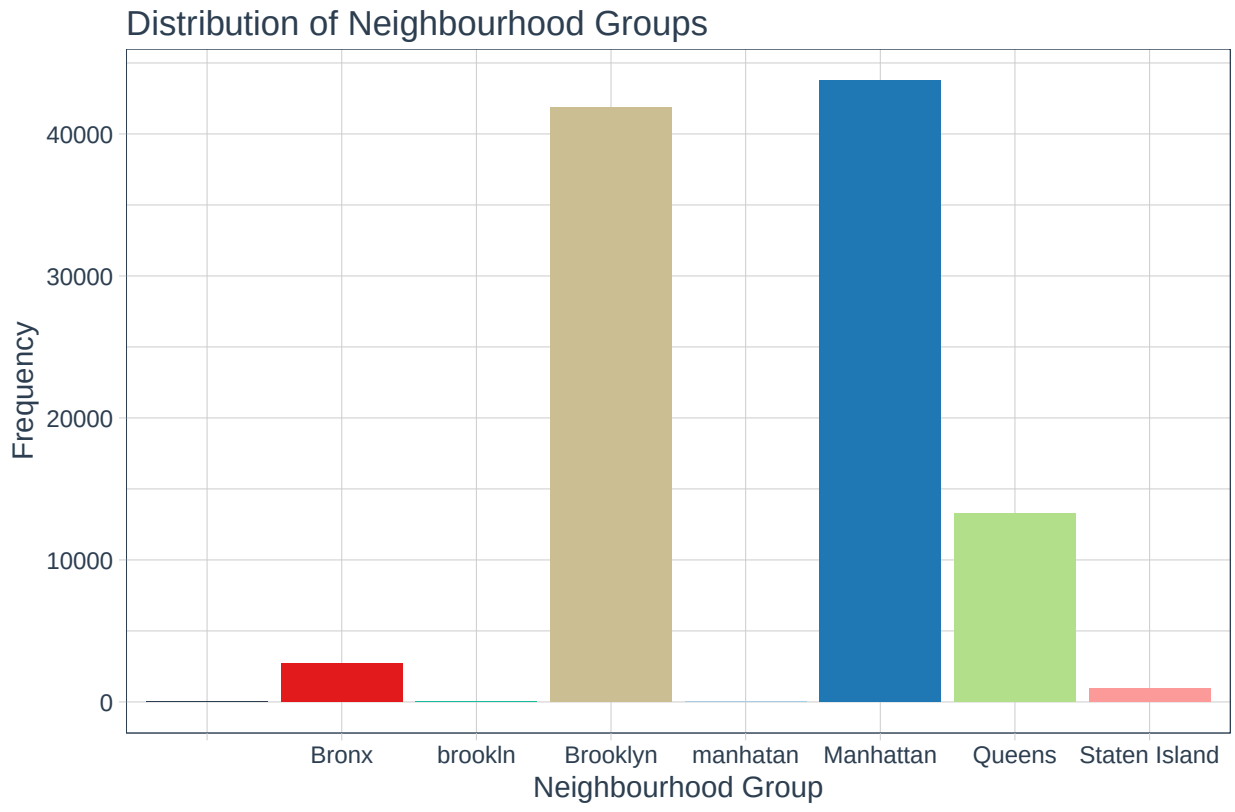
```
##   neighbourhood.group     n percentage
## 1                          29     0.028%
## 2               Bronx  2712     2.643%
## 3            Brooklyn 41842    40.782%
## 4           Manhattan 43792    42.683%
## 5              Queens 13267    12.931%
## 6       Staten Island   955     0.931%
## 7             brookln     1     0.001%
```

```
## 8            manhatan     1      0.001%
```

```
#graph 1
airbnb %>%
  ggplot() +
  geom_bar(mapping = aes(x = neighbourhood.group, fill = neighbourhood.group)) +
  scale_fill_tq() +

    labs( x = "Neighbourhood Group",
          y = "Frequency",
          title = "Distribution of Neighbourhood Groups",
        caption = "Manhattan seems to have the most airbnbs listed") +

  theme_tq() +
  theme(legend.position = "none")
```



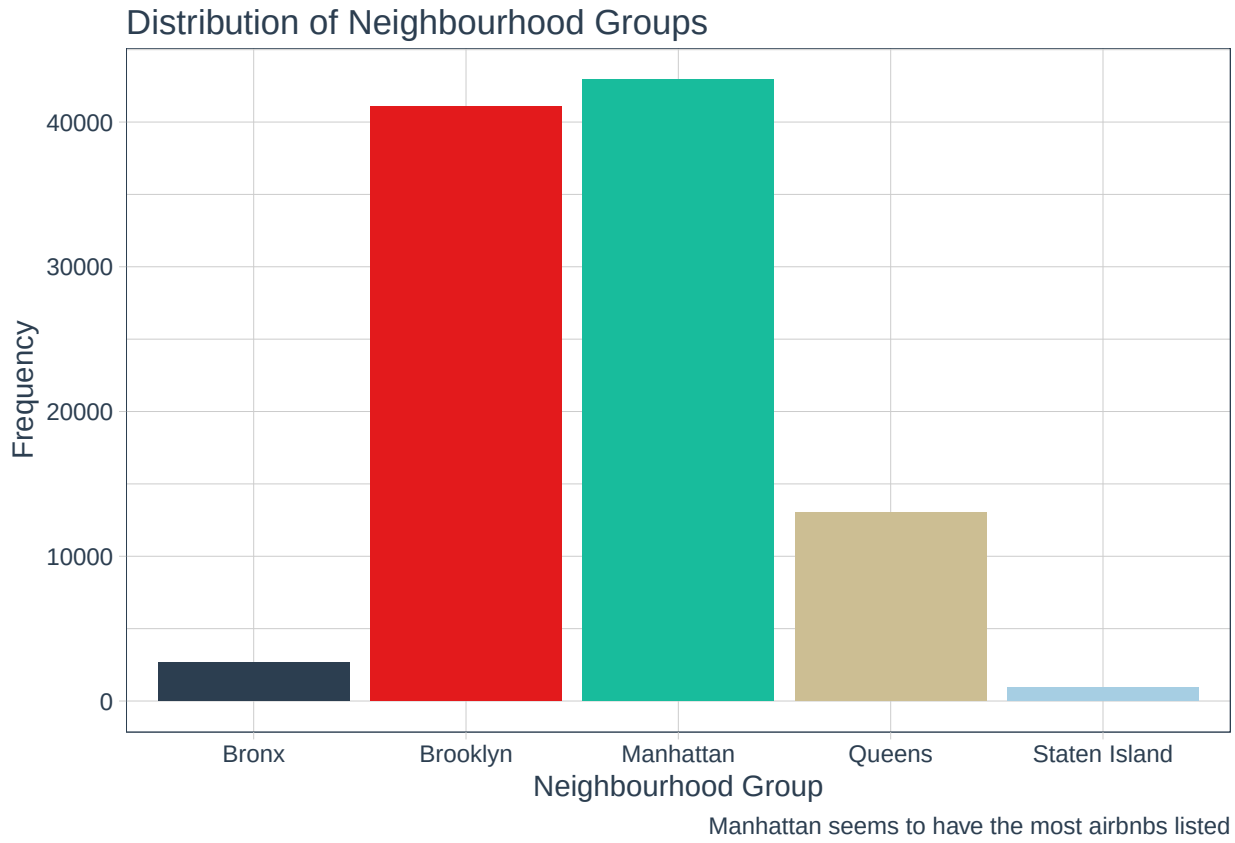

```
#getting rid of miss spellings and blank answers
airbnb_v_c <- airbnb %>%
  mutate(neighbourhood.group = case_when(
    neighbourhood.group == "Bronx" ~ "Bronx",
    neighbourhood.group == "Brooklyn" ~ "Brooklyn",
    neighbourhood.group == "brookln" ~ "Brooklyn",
    neighbourhood.group == "manhatan" ~ "Manhattan",
    neighbourhood.group == "Manhattan" ~ "Manhattan",
    neighbourhood.group == "Queens" ~ "Queens",
```

```
    neighbourhood.group == "Staten Island" ~ "Staten Island",
    TRUE ~ NA_character_
  ))

#graph 2
airbnb_v_c %>% drop_na %>%
  ggplot() +
  geom_bar(mapping = aes(x = neighbourhood.group, fill = neighbourhood.group)) +
  scale_fill_tq() +

    labs( x = "Neighbourhood Group",
          y = "Frequency",
          title = "Distribution of Neighbourhood Groups",
        caption = "Manhattan seems to have the most airbnbs listed") +

  theme_tq() +
  theme(legend.position = "none")
```



There seems to be some missing values that people either did not fill out or forgot. Also someone people have miss spelled neighborhood groups so it made it there own category. You can also see this on the graph where the miss spelled data is.

Room Type

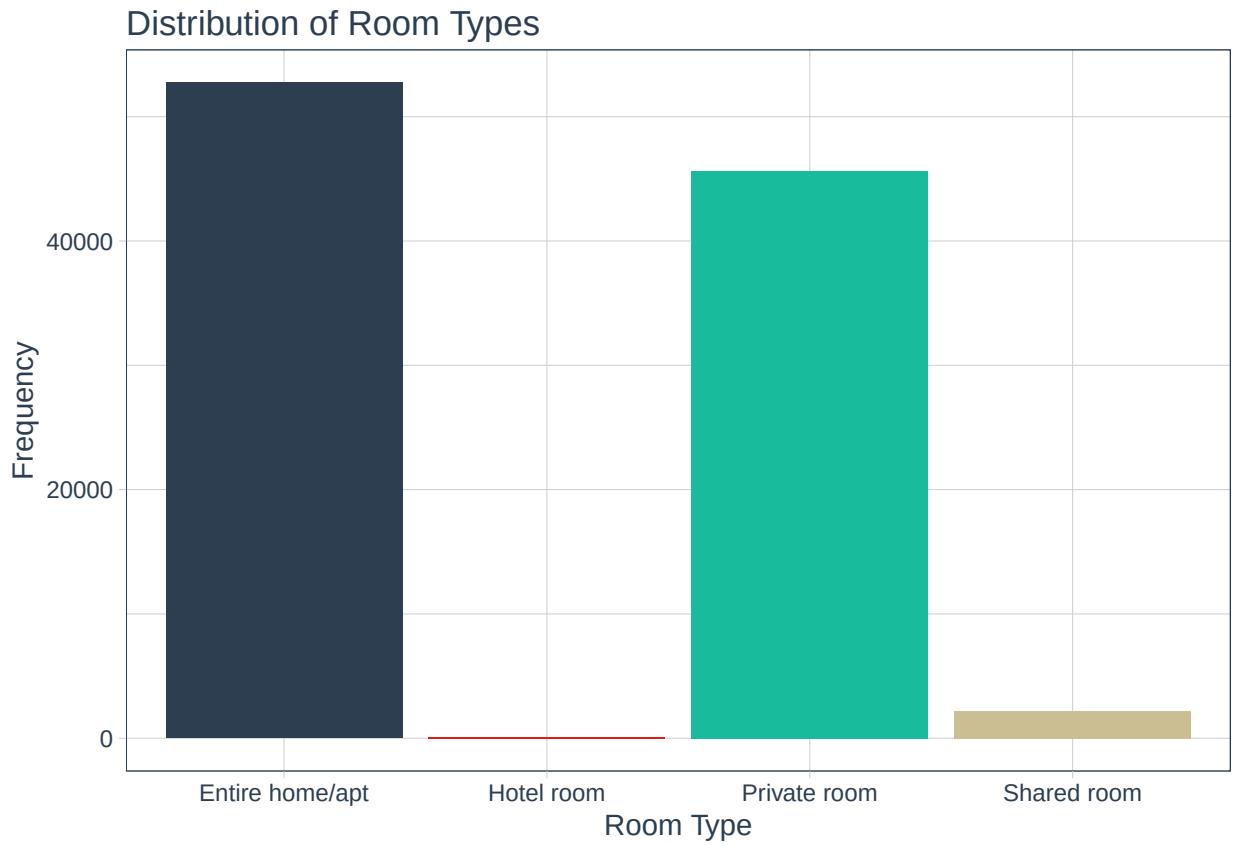
```
airbnb %>%
  count(room.type) %>%
  mutate(percentage = (n/sum(n)) %>% scales::percent()) %>%
  arrange(desc(n))
```

```
##         room.type     n percentage
## 1 Entire home/apt 53701      52.3%
## 2    Private room 46556      45.4%
## 3     Shared room  2226       2.2%
## 4      Hotel room   116       0.1%
```

```
airbnb_v_c %>% drop_na %>%
  ggplot() +
  geom_bar(mapping = aes(x = room.type, fill = room.type)) +
  scale_fill_tq() +

    labs( x = "Room Type",
          y = "Frequency",
          title = "Distribution of Room Types") +

  theme_tq() +
  theme(legend.position = "none")
```



It seems the majority of room types are entire home/apartment and private room.

-

Section 2: Now I will be checking the relationship between the variables

–

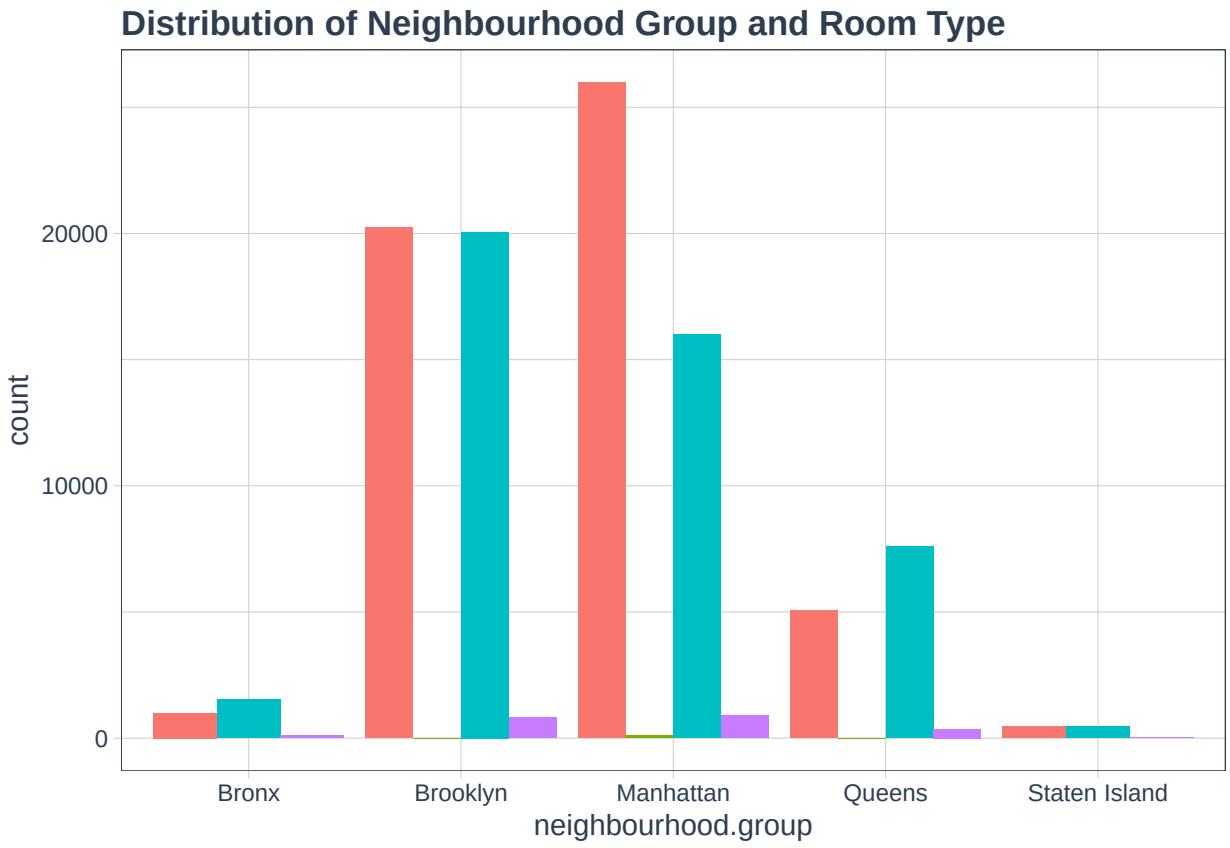
I: Two Categorical Variables

```
airbnb_v_c %>% drop_na() %>%
  count(room.type,neighbourhood.group) %>%
  mutate(percentage = (n/sum(n)) %>% scales::percent()) %>%
  arrange(desc(n))
```

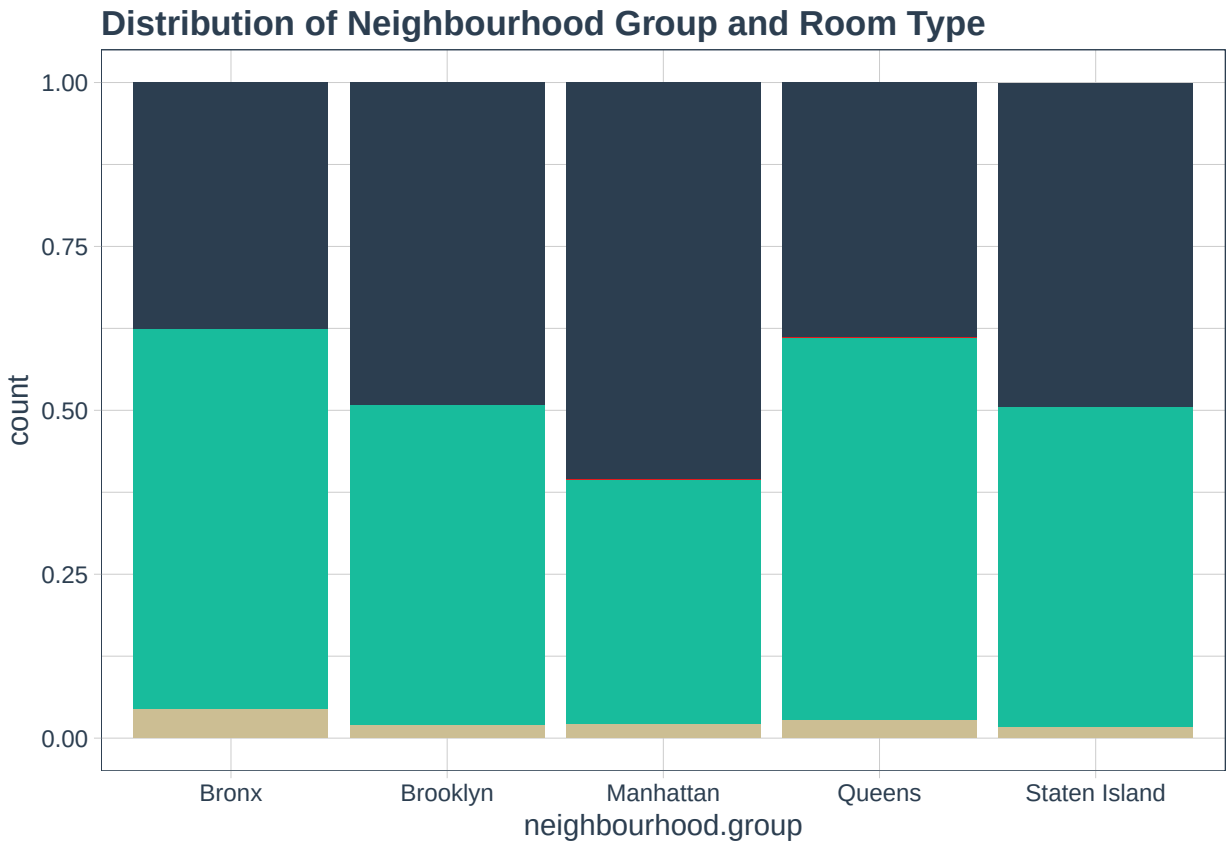
Room Type and Neighbourhood Group

```
##          room.type neighbourhood.group     n percentage
## 1  Entire home/apt           Manhattan 25983   25.8096%
## 2  Entire home/apt            Brooklyn 20224   20.0890%
## 3     Private room            Brooklyn 20063   19.9291%
## 4     Private room           Manhattan 15987   15.8803%
## 5     Private room              Queens  7584    7.5334%
## 6  Entire home/apt              Queens  5065    5.0312%
## 7     Private room               Bronx  1543    1.5327%
## 8  Entire home/apt               Bronx   999    0.9923%
## 9      Shared room           Manhattan   896    0.8900%
## 10     Shared room            Brooklyn   806    0.8006%
## 11 Entire home/apt       Staten Island   465    0.4619%
## 12    Private room       Staten Island   458    0.4549%
## 13     Shared room              Queens   357    0.3546%
## 14     Shared room               Bronx   115    0.1142%
## 15      Hotel room           Manhattan    96    0.0954%
## 16     Shared room       Staten Island    15    0.0149%
## 17      Hotel room            Brooklyn     8    0.0079%
## 18      Hotel room              Queens     8    0.0079%
```

```
#graph 1
airbnb_v_c %>% drop_na() %>%
  ggplot(mapping = aes(x = neighbourhood.group, fill = room.type)) +
  geom_bar(position = "dodge") +
  theme_tq() +
  labs(title = "Distribution of Neighbourhood Group and Room Type") +
   theme(legend.position = "none",
        plot.title = element_text(size = 13, face = "bold"))
```




```
#graph 2
airbnb_v_c %>% drop_na() %>%
  ggplot(mapping = aes(x = neighbourhood.group, fill = room.type))+
  geom_bar(position = "fill") +
  scale_fill_tq() +
  theme_tq() +
  labs(title = "Distribution of Neighbourhood Group and Room Type") +
  theme(legend.position = "none",
        plot.title = element_text(size = 13, face = "bold"))
```



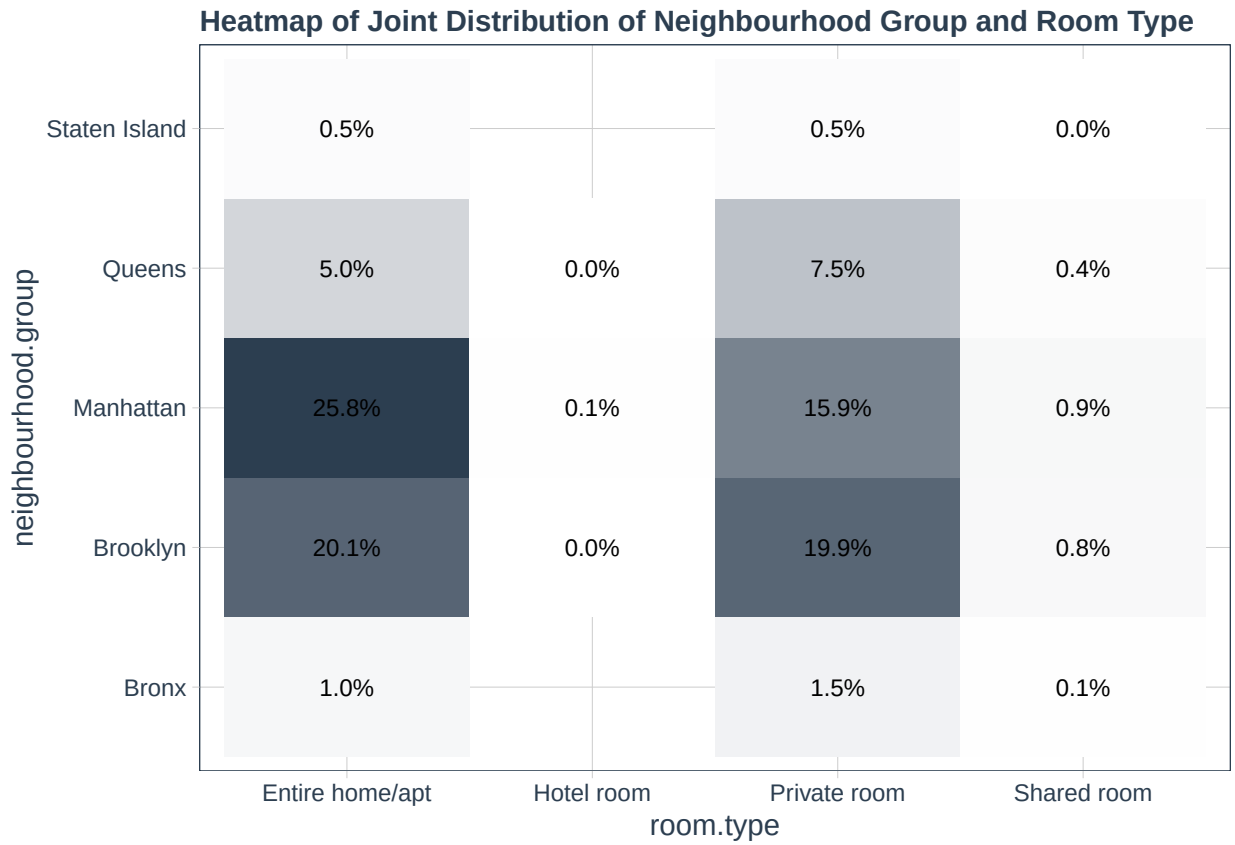

```
joint_mb <- airbnb_v_c %>% drop_na() %>%
  count(neighbourhood.group,room.type) %>%
  mutate(pct = n/sum(n))

#graph 3
joint_mb %>%
  ggplot(mapping = aes(x = room.type, y = neighbourhood.group)) +
  geom_tile(mapping = aes(fill = pct)) +
  geom_text(aes(label = pct %>% scales::percent(accuracy = 0.1)), size = 3) +

  #labels and scales

  scale_fill_gradient(low = "white",high = palette_light()[1]) +#palette_light() shows color
  labs(title = "Heatmap of Joint Distribution of Neighbourhood Group and Room Type")+

  #themes
  theme_tq() +
  theme(legend.position = "none",
        plot.title = element_text(size = 11, face = "bold"),
        plot.caption = element_text(face = "bold.italic"))
```



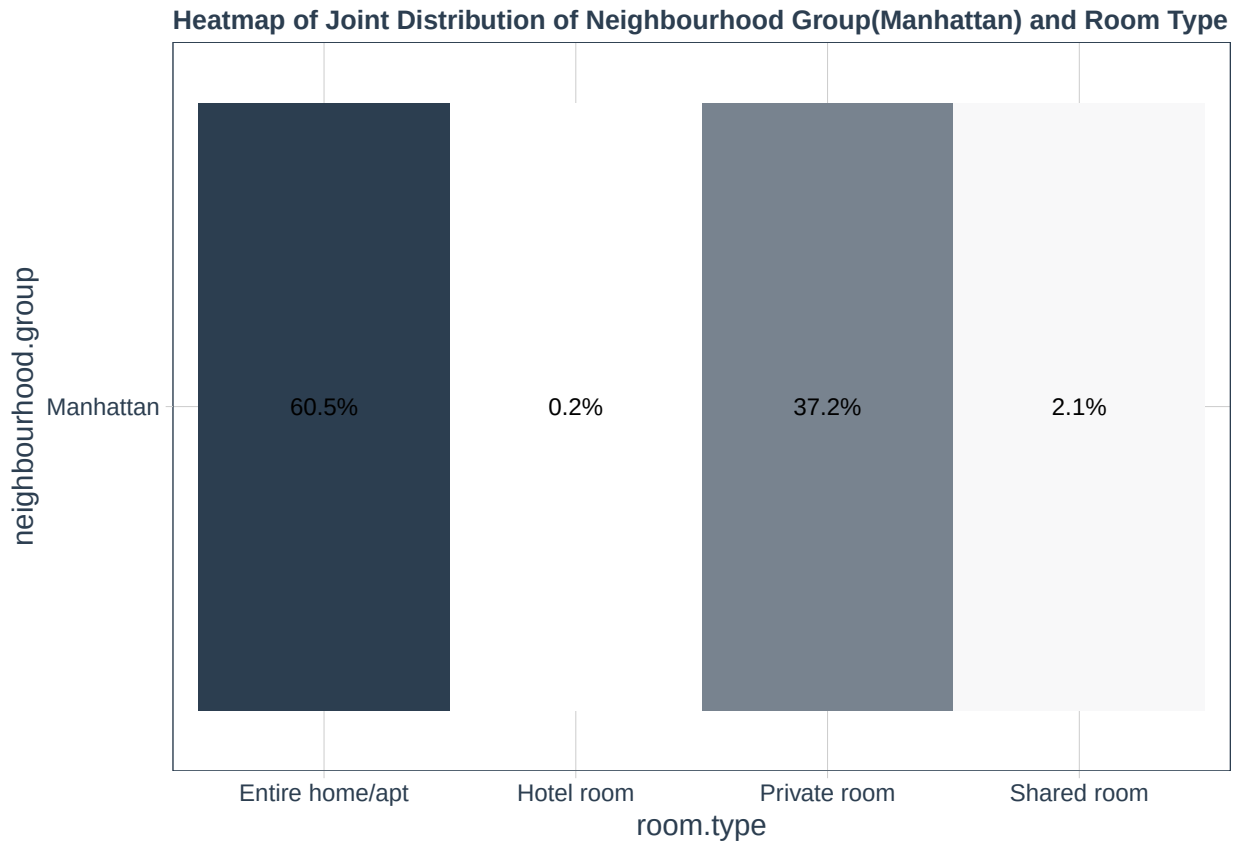

```
#graph 4
joint_mb <- airbnb_v_c %>% drop_na() %>%
  filter(neighbourhood.group == "Manhattan") %>%
  count(neighbourhood.group,room.type) %>%
  mutate(pct = n/sum(n))

joint_mb %>%
  ggplot(mapping = aes(x = room.type, y = neighbourhood.group)) +
  geom_tile(mapping = aes(fill = pct)) +
  geom_text(aes(label = pct %>% scales::percent(accuracy = 0.1)), size = 3) +

  #labels and scales

  scale_fill_gradient(low = "white",high = palette_light()[1]) +#palette_light() shows color
  labs(title = "Heatmap of Joint Distribution of Neighbourhood Group(Manhattan) and Room Type")+

  #themes
  theme_tq() +
  theme(legend.position = "none",
        plot.title = element_text(size = 10, face = "bold"),
        plot.caption = element_text(face = "bold.italic"))
```



-

II: Two continuous variables

```
cor_p_s = cor(airbnb$price,airbnb$service.fee, use = "complete.obs")

cor_p_s
```

Price and Service Fees

```
## [1] 0.9999909
```

```
airbnbpsf <- airbnb %>% drop_na() %>%
  count(price, service.fee) %>%
  arrange(desc(n))
head(airbnbpsf)
```

```
##   price service.fee   n
## 1   206          41 131
## 2   833         167 128
## 3  1056         211 128
```

```
## 4   481          96 127
## 5   573         115 126
## 6   972         194 123
```

```
airbnb %>% drop_na() %>%
  ggplot(mapping = aes(x = price, y = service.fee)) +
  geom_point(alpha = 0.1) +
  theme_tq() +
  labs(title = "Distribution of Service Price and Price") +
  theme(legend.position = "none",
        plot.title = element_text(size = 15, face = "bold"))
```

-

III: One Categorical Variable and One Continuous Variable

```
airbnb_v_c <- airbnb_v_c %>% drop_na() %>%
  count(room.type, number.of.reviews) %>%
  arrange(desc(number.of.reviews))
head(airbnb_v_c)
```

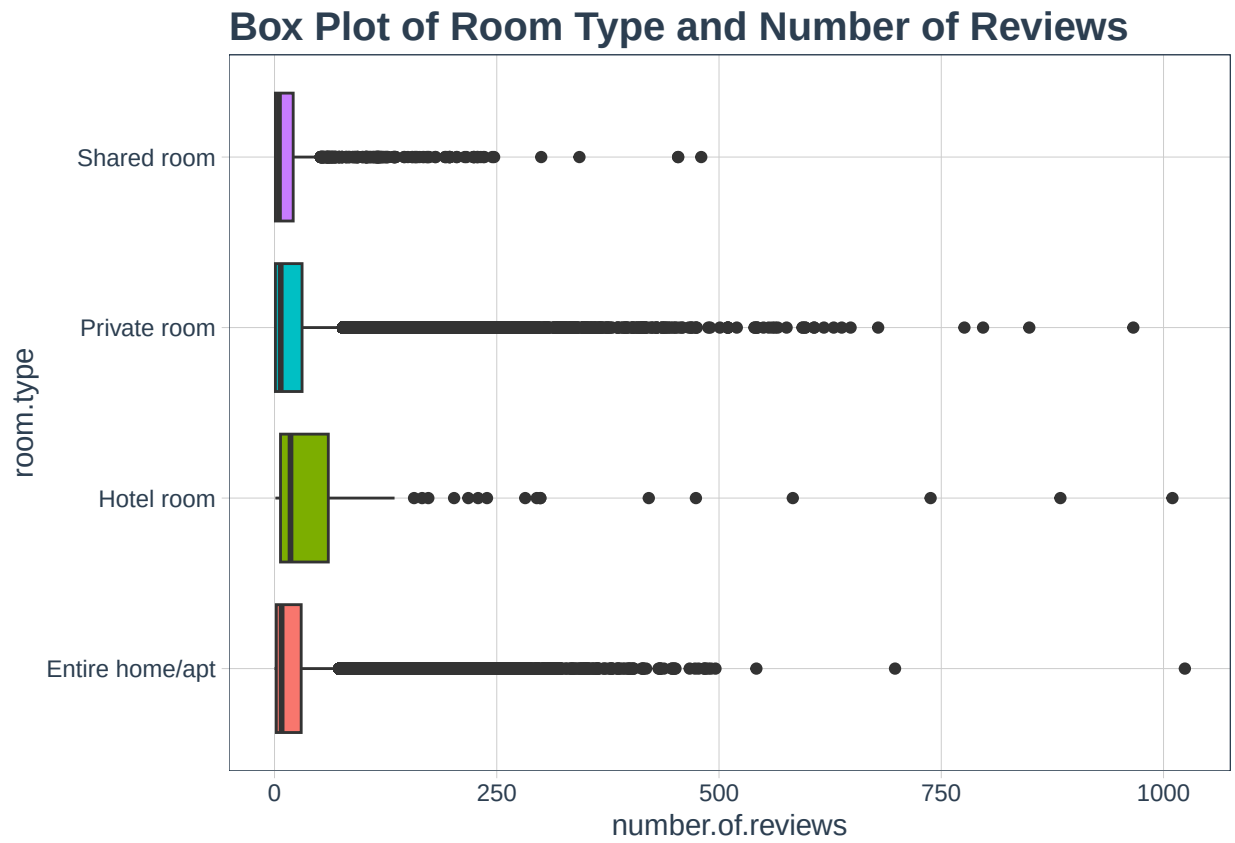
Room Type and Number of Reviews

```
##         room.type number.of.reviews n
## 1 Entire home/apt              1024 1
## 2      Hotel room              1010 1
## 3    Private room               966 1
## 4      Hotel room               884 1
## 5    Private room               849 1
## 6    Private room               797 1
```

```
airbnb %>% drop_na() %>%
  group_by(room.type) %>%
  summarise(avg_number_reviews = mean(number.of.reviews))
```

```
## # A tibble: 4 x 2
##   room.type       avg_number_reviews
##   <chr>                        <dbl>
## 1 Entire home/apt               27.2
## 2 Hotel room                    80.2
## 3 Private room                  27.6
## 4 Shared room                   20.5
```

```
airbnb %>% drop_na() %>%
  ggplot(mapping = aes(x = room.type , y = number.of.reviews ,fill = room.type)) +
  geom_boxplot() +
  theme_tq() +
  scale_color_tq()+
  coord_flip() +
  labs(title = "Box Plot of Room Type and Number of Reviews") +
  theme(legend.position = "none",
        plot.title = element_text(size = 15, face = "bold"))
```



-

IV: Two continuous variables and One Categorical Variable

```
airbnb_ccc <- airbnb %>% drop_na() %>%
  filter(number.of.reviews > 0) %>%
  count(number.of.reviews,price,room.type) %>%
  arrange(desc(number.of.reviews))
 head(airbnb_ccc)
```

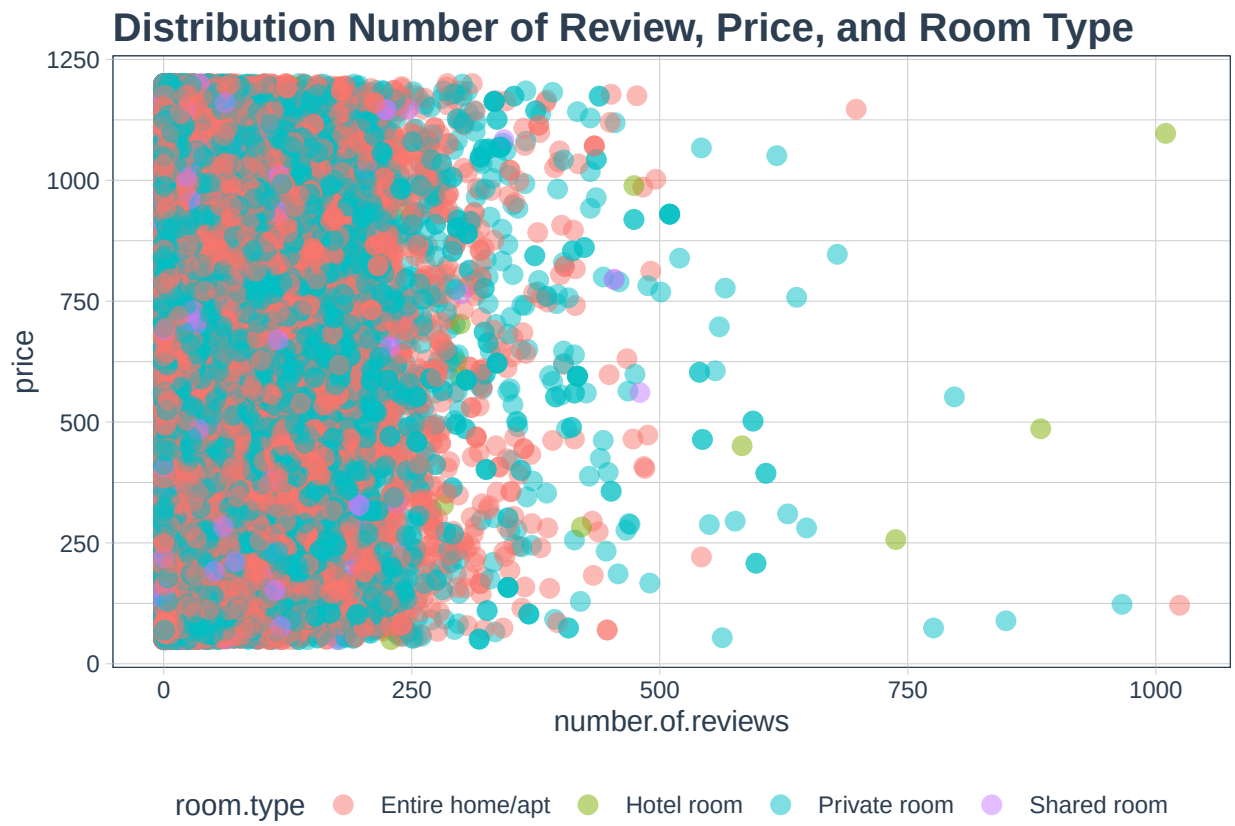
Number of Review, Price, and Room Type

```
##   number.of.reviews price       room.type n
## 1              1024   121 Entire home/apt 1
## 2              1010  1097      Hotel room 1
## 3               966   123    Private room 1
## 4               884   486      Hotel room 1
## 5               849    89    Private room 1
## 6               797   552    Private room 1
```

```
airbnb_ccc %>%
  group_by(room.type) %>%
  summarise(sum_rev = sum(number.of.reviews))
```

```
## # A tibble: 4 x 2
##   room.type       sum_rev
##   <chr>             <int>
## 1 Entire home/apt  957856
## 2 Hotel room         8971
## 3 Private room     795385
## 4 Shared room       28018
```

```
airbnb %>% drop_na() %>%
 ggplot(mapping = aes(x = number.of.reviews, y = price , color = room.type)) +
  geom_point(size = 3, alpha = .5) +
  theme_tq() +
  labs(title = "Distribution Number of Review, Price, and Room Type") +
  theme(
        plot.title = element_text(size = 15, face = "bold"))
```



Chapter 3: Research Questions and findings from Section 2

In section 2 of chapter 2 I was looking at two categorical variables, two continuous variables, One categorical variable and one continuous variable, and two continuous variables and one categorical variable. The variables

that I picked for two categorical variables were room type and neighbourhood group. I first found the analytical percentage distribution. For the first visual graph I decided to use a bar graph. I noticed that Manhattan was the only one with hotel room types. Then I did a bar graph but I used the fill option to give another look. I also could still only see hotel room types for Manhattan. I decided to look at the joint distribution. I saw that Manhattan had the most so I did another joint distribution of just Manhattan to see the breakdown. I saw that 60% was the entire home/apartment. For the two continuous variables I picked price and service fee. Since both were analytical I did a correlation. To my surprise it was .99. So I checked with a scatter plot to see if it was a .99 and it indeed gave me a nearly perfect positive linear line. The next section was one continuous and one categorical. I picked a number of reviews and room type. So I found the combination of number of reviews and room type. The highest number of combinations was zero reviews which is not surprising because the average person does not liking leaving reviews on anything. So I decided to arrange in descending order by total number of reviews. I saw that the entire home/apartment and hotel had the highest amount of reviews. Then I decided to find the average number of reviews by room type. I was shocked to find the hotel had the highest amount of average reviews. I think the reason could be that some hotels might have incentives to leave reviews. It is also surprising because hotel has the least amount of room types. So I thought using a box plot would look for the amount of reviews against the room types. You can also see here that the hotel does not have as many reviews compared to the other ones. Since the number of reviews for the other types besides hotel is much more dense on the lower end of number reviews this might be why the average number of reviews is much lower compared to hotel. For the last section two continuous variables and one categorical variable. I picked price, number of reviews, and room type. I checked the combination of the three and of course 0 reviews was the highest combination. SO I arranged by the largest amount of reviews which was still hotel and entire home/apartment. The interesting thing is that price between hotel and entire home/apartment was very large. I also decided to check who had the most reviews. I found the summation of reviews for room type. So when graphing I used a scatter plot. I made the x and y the continuous and the color categorical. When looking at the graph you can see that majority of the reviews are entire home/apartment.