AIRBNB DATASET: EXPLORATORY DATA ANALYSIS

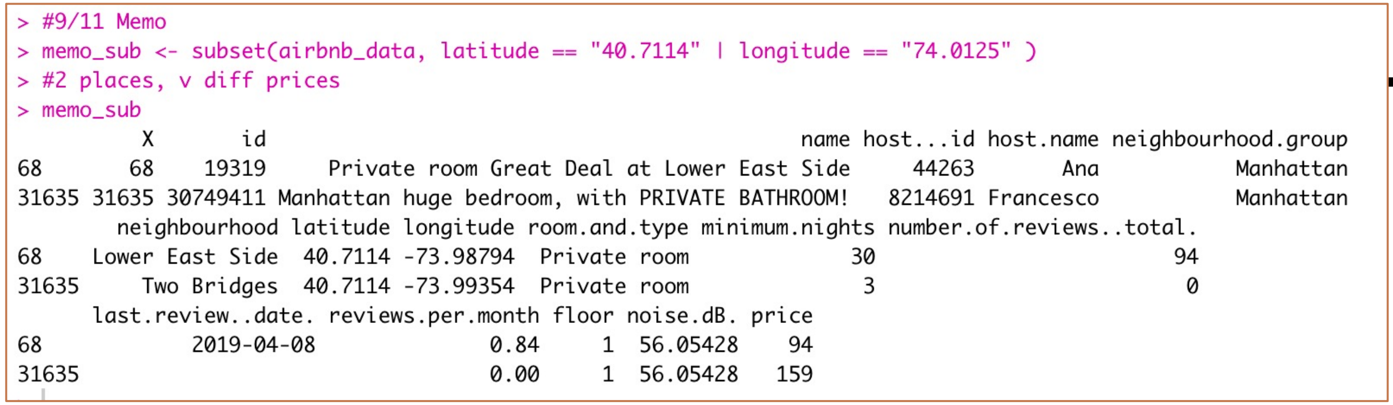
Presentation by:

Muskan Burman

CS:310: Data Science Capstone

POINTS OF INTEREST – SOME OUTLIERS IN DATA

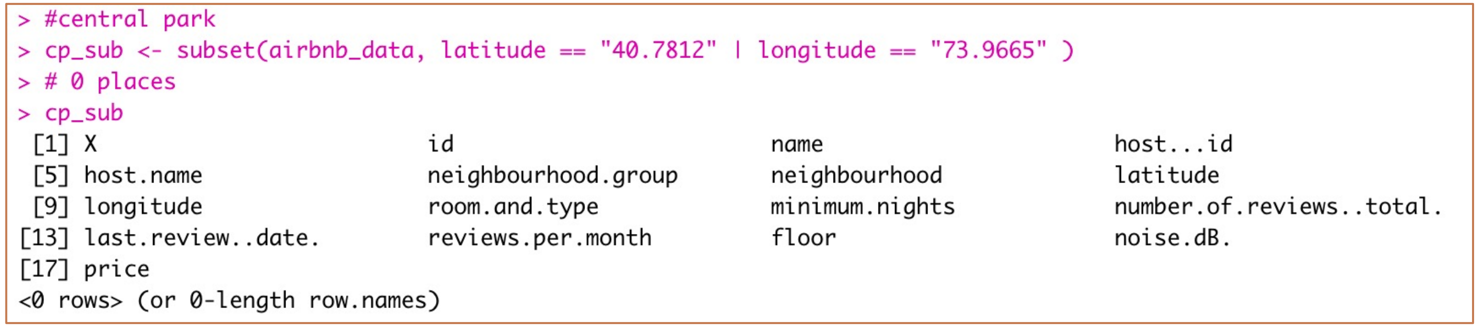
- I looked at various points of interest – places in NYC that are popular tourist spots to see if their location affected the prices of Airbnbs in that area. Although completely different, each POI was similar from the aspect of not being a very popular Airbnb spot, as there weren’t many rooms in the dataset for either of them.
- I looked at Airbnbs near the 9/11 memorial, Central Park, Empire State Building, Statue of Liberty, Madison Square Garden, and the JFK Airport.



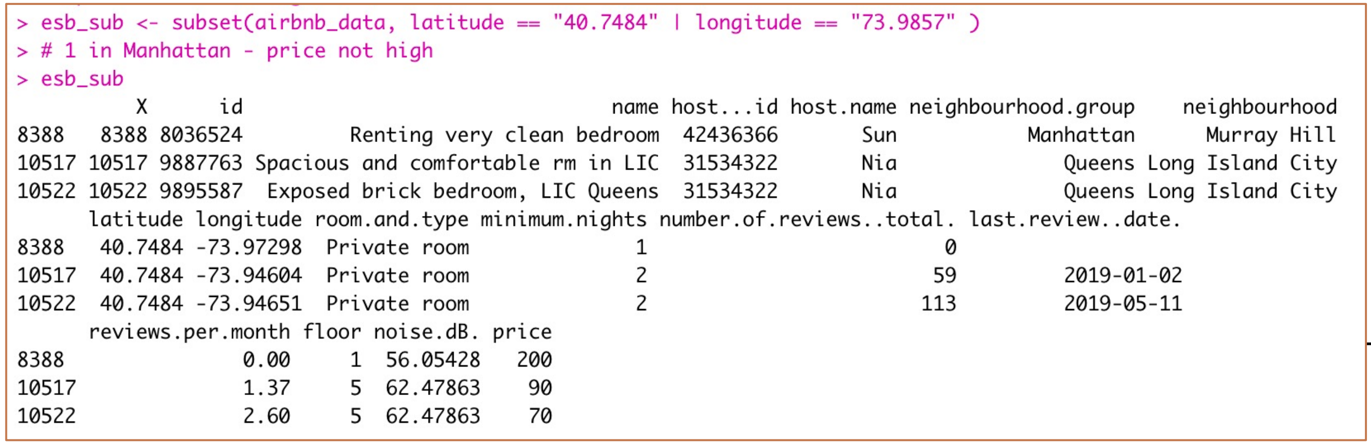
```
> #9/11 Memo
> memo_sub <- subset(airbnb_data, latitude == "40.7114" | longitude == "74.0125" )
> #2 places, v diff prices
> memo_sub
          X       id                                       name host...id host.name neighbourhood.group
68       68    19319      Private room Great Deal at Lower East Side     44263       Ana           Manhattan
31635 31635 30749411 Manhattan huge bedroom, with PRIVATE BATHROOM!   8214691 Francesco           Manhattan
        neighbourhood latitude longitude room.and.type minimum.nights number.of.reviews..total.
68    Lower East Side  40.7114 -73.98794  Private room             30                        94
31635     Two Bridges  40.7114 -73.99354  Private room              3                         0
      last.review..date. reviews.per.month floor noise.dB. price
68            2019-04-08              0.84     1  56.05428    94
31635                                 0.00     1  56.05428   159
```

Only 2 private rooms could be found near the 9/11 memorial, and their prices varied significantly.

As we can see here, no room near the central park could be found in the dataset.

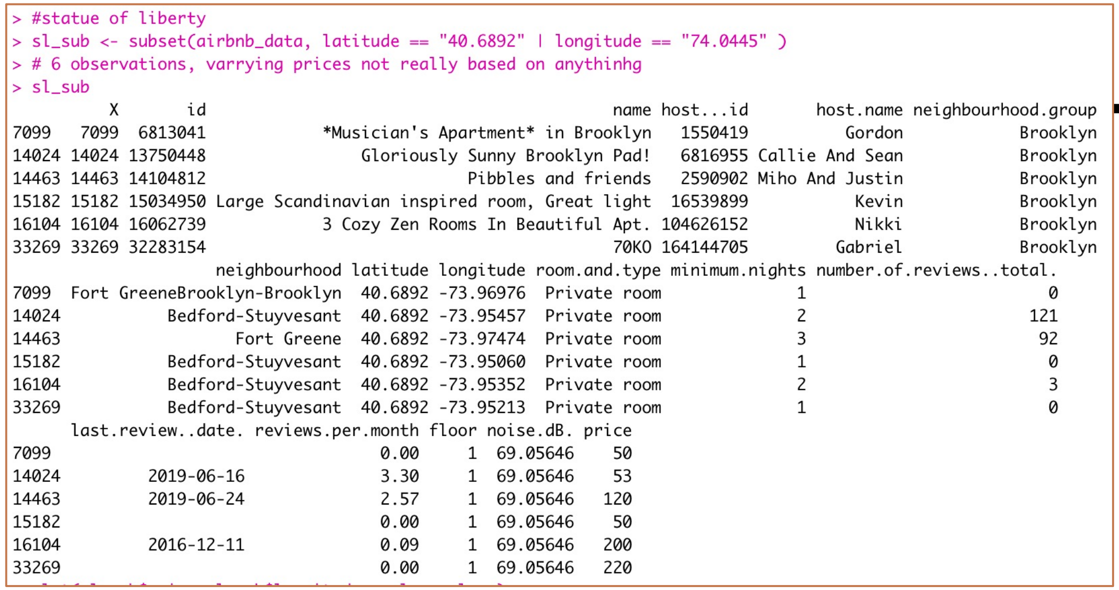


```
> #central park
> cp_sub <- subset(airbnb_data, latitude == "40.7812" | longitude == "73.9665" )
> # 0 places
> cp_sub
 [1] X                          id                         name                       host...id
 [5] host.name                  neighbourhood.group        neighbourhood              latitude
 [9] longitude                  room.and.type              minimum.nights             number.of.reviews..total.
[13] last.review..date.         reviews.per.month          floor                      noise.dB.
[17] price
<0 rows> (or 0-length row.names)
```



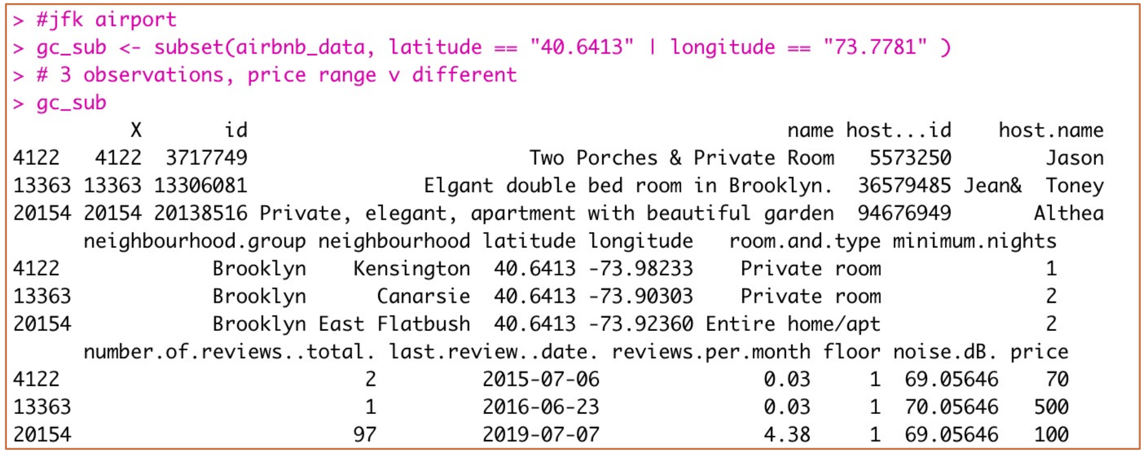
```
> esb_sub <- subset(airbnb_data, latitude == "40.7484" | longitude == "73.9857" )
> # 1 in Manhattan - price not high
> esb_sub
          X      id                           name host...id host.name neighbourhood.group     neighbourhood
8388   8388 8036524          Renting very clean bedroom  42436366       Sun           Manhattan       Murray Hill
10517 10517 9887763 Spacious and comfortable rm in LIC  31534322       Nia              Queens Long Island City
10522 10522 9895587  Exposed brick bedroom, LIC Queens  31534322       Nia              Queens Long Island City
      latitude longitude room.and.type minimum.nights number.of.reviews..total. last.review..date.
8388   40.7484 -73.97298  Private room              1                         0
10517  40.7484 -73.94604  Private room              2                        59         2019-01-02
10522  40.7484 -73.94651  Private room              2                       113         2019-05-11
      reviews.per.month floor noise.dB. price
8388               0.00     1  56.05428   200
10517              1.37     5  62.47863    90
10522              2.60     5  62.47863    70
```

There was only one private room available near the Empire State Building in Manhattan.



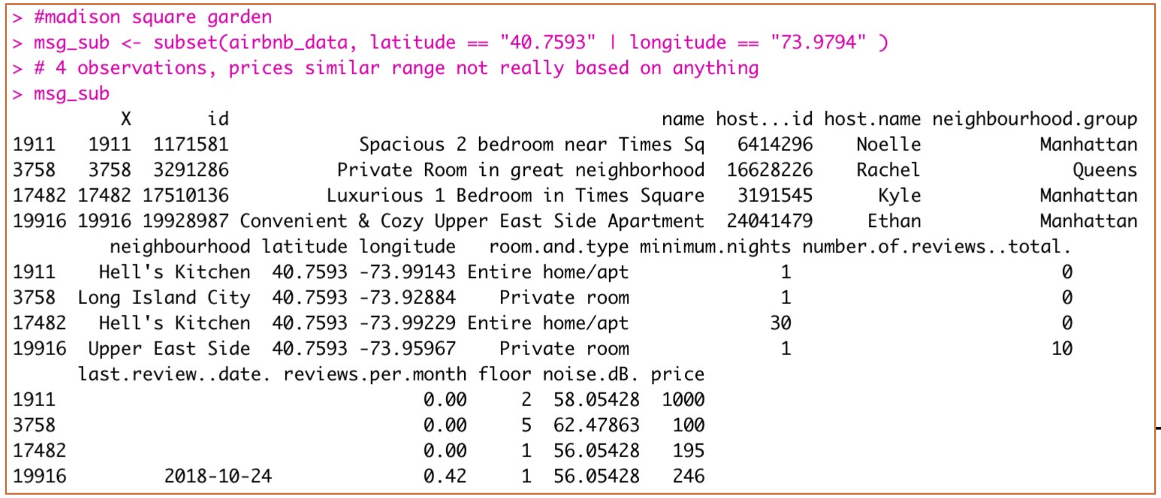
```
> #statue of liberty
> sl_sub <- subset(airbnb_data, latitude == "40.6892" | longitude == "74.0445" )
> # 6 observations, varrying prices not really based on anythinhg
> sl_sub
          X       id                                   name host...id       host.name neighbourhood.group
7099   7099  6813041           *Musician's Apartment* in Brooklyn   1550419          Gordon            Brooklyn
14024 14024 13750448             Gloriously Sunny Brooklyn Pad!   6816955 Callie And Sean            Brooklyn
14463 14463 14104812                        Pibbles and friends   2590902 Miho And Justin            Brooklyn
15182 15182 15034950 Large Scandinavian inspired room, Great light  16539899           Kevin            Brooklyn
16104 16104 16062739             3 Cozy Zen Rooms In Beautiful Apt. 104626152           Nikki            Brooklyn
33269 33269 32283154                                          70KO 164144705         Gabriel            Brooklyn
                     neighbourhood latitude longitude room.and.type minimum.nights number.of.reviews..total.
7099  Fort GreeneBrooklyn-Brooklyn  40.6892 -73.96976  Private room              1                         0
14024           Bedford-Stuyvesant  40.6892 -73.95457  Private room              2                       121
14463                  Fort Greene  40.6892 -73.97474  Private room              3                        92
15182           Bedford-Stuyvesant  40.6892 -73.95060  Private room              1                         0
16104           Bedford-Stuyvesant  40.6892 -73.95352  Private room              2                         3
33269           Bedford-Stuyvesant  40.6892 -73.95213  Private room              1                         0
      last.review..date. reviews.per.month floor noise.dB. price
7099                                  0.00     1  69.05646    50
14024         2019-06-16              3.30     1  69.05646    53
14463         2019-06-24              2.57     1  69.05646   120
15182                                 0.00     1  69.05646    50
16104         2016-12-11              0.09     1  69.05646   200
33269                                 0.00     1  69.05646   220
```

Statue of Liberty has 6 places near rooms in its neighbourhood, all private. Their prices also varied significantly but were almost independent of the other factors.



```
> #jfk airport
> gc_sub <- subset(airbnb_data, latitude == "40.6413" | longitude == "73.7781" )
> # 3 observations, price range v different
> gc_sub
          X       id                                           name host...id    host.name
4122   4122  3717749                          Two Porches & Private Room   5573250        Jason
13363 13363 13306081              Elgant double bed room in Brooklyn.  36579485 Jean&  Toney
20154 20154 20138516 Private, elegant, apartment with beautiful garden  94676949       Althea
      neighbourhood.group neighbourhood latitude longitude   room.and.type minimum.nights
4122             Brooklyn    Kensington  40.6413 -73.98233    Private room              1
13363            Brooklyn      Canarsie  40.6413 -73.90303    Private room              2
20154            Brooklyn East Flatbush  40.6413 -73.92360 Entire home/apt              2
      number.of.reviews..total. last.review..date. reviews.per.month floor noise.dB. price
4122                          2         2015-07-06              0.03     1  69.05646    70
13363                         1         2016-06-23              0.03     1  70.05646   500
20154                        97         2019-07-07              4.38     1  69.05646   100
```

Again, only 3 rooms could be found near the JFK airport, with 2 of them being private and the third one being an entire house. Again, the difference in price and their range was significant.

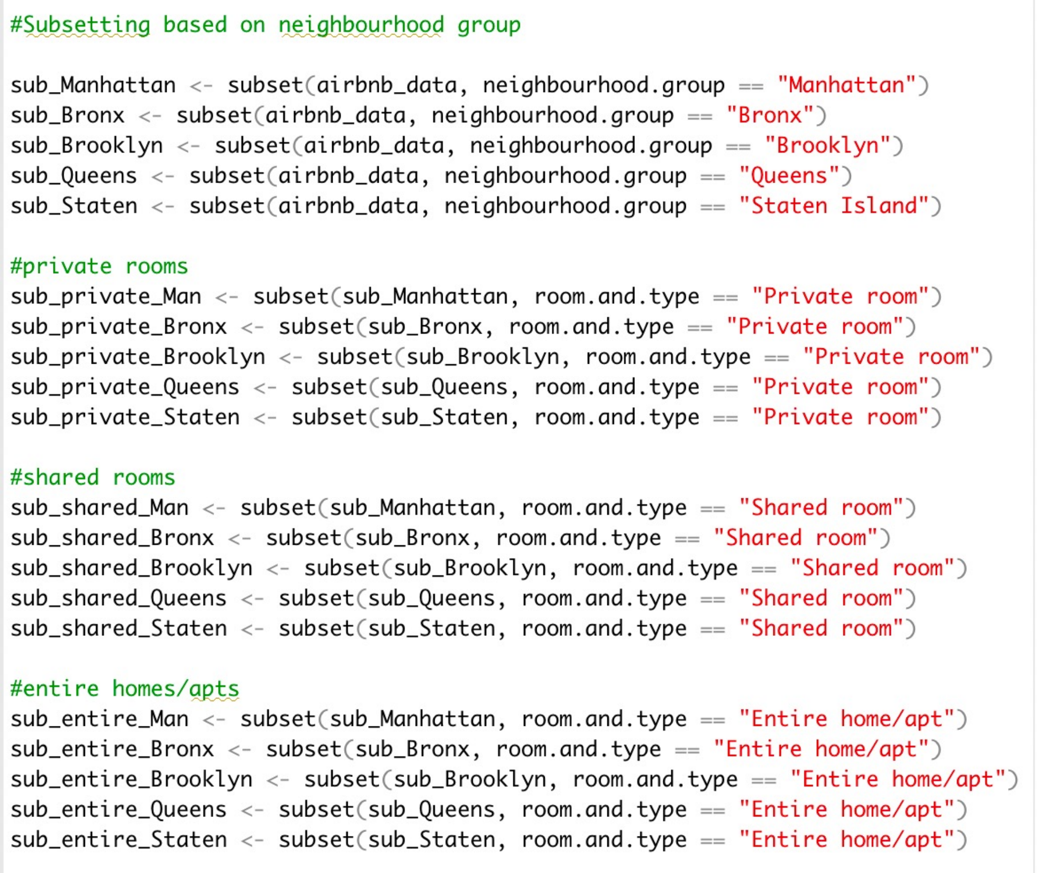


```
> #madison square garden
> msg_sub <- subset(airbnb_data, latitude == "40.7593" | longitude == "73.9794" )
> # 4 observations, prices similar range not really based on anything
> msg_sub
          X       id                               name host...id host.name neighbourhood.group
1911   1911  1171581         Spacious 2 bedroom near Times Sq   6414296    Noelle           Manhattan
3758   3758  3291286        Private Room in great neighborhood  16628226    Rachel              Queens
17482 17482 17510136          Luxurious 1 Bedroom in Times Square   3191545      Kyle           Manhattan
19916 19916 19928987 Convenient & Cozy Upper East Side Apartment  24041479     Ethan           Manhattan
         neighbourhood latitude longitude   room.and.type minimum.nights number.of.reviews..total.
1911    Hell's Kitchen  40.7593 -73.99143 Entire home/apt              1                         0
3758  Long Island City  40.7593 -73.92884    Private room              1                         0
17482   Hell's Kitchen  40.7593 -73.99229 Entire home/apt             30                         0
19916  Upper East Side  40.7593 -73.95967    Private room              1                        10
      last.review..date. reviews.per.month floor noise.dB. price
1911                                  0.00     2  58.05428  1000
3758                                  0.00     5  62.47863   100
17482                                 0.00     1  56.05428   195
19916         2018-10-24              0.42     1  56.05428   246
```

Madison Square Garden had 3 rooms in its neighbourhood, 2 being private while the other one being an entire house. Price range was significantly different for these observations as well.

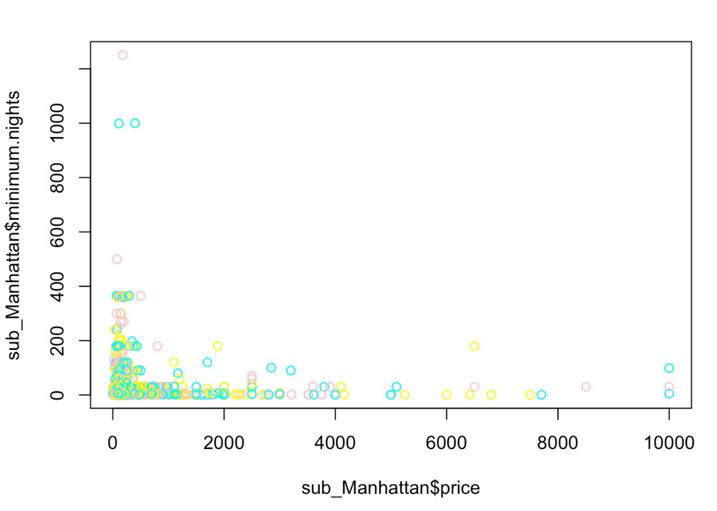
SUBSETTING THE DATASET

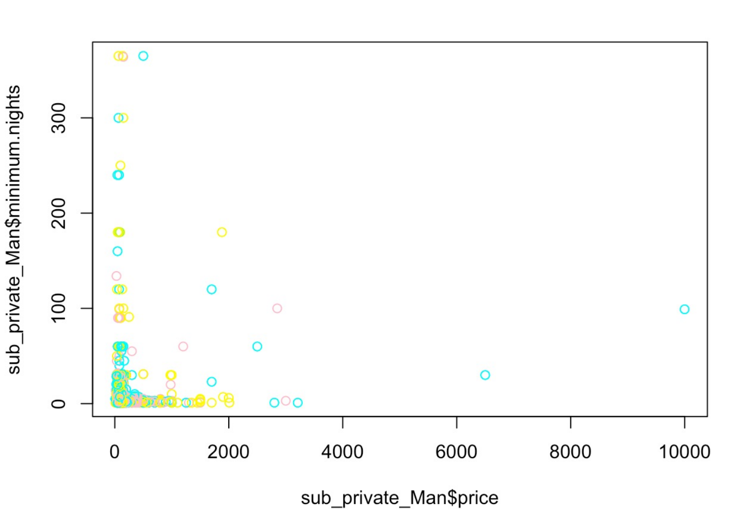
To look at the relationships of the various attributes with price, I subsetted the dataset first based on the neighbourhood group, resulting in 5 different sets, and then further based on the room type, resulting in 15 more sets to get a total of 20. I then used these subsets, as required, to identify the factors that affected the price of an Airbnb room.



```
#Subsetting based on neighbourhood group

sub_Manhattan <- subset(airbnb_data, neighbourhood.group == "Manhattan")
sub_Bronx <- subset(airbnb_data, neighbourhood.group == "Bronx")
sub_Brooklyn <- subset(airbnb_data, neighbourhood.group == "Brooklyn")
sub_Queens <- subset(airbnb_data, neighbourhood.group == "Queens")
sub_Staten <- subset(airbnb_data, neighbourhood.group == "Staten Island")

#private rooms
sub_private_Man <- subset(sub_Manhattan, room.and.type == "Private room")
sub_private_Bronx <- subset(sub_Bronx, room.and.type == "Private room")
sub_private_Brooklyn <- subset(sub_Brooklyn, room.and.type == "Private room")
sub_private_Queens <- subset(sub_Queens, room.and.type == "Private room")
sub_private_Staten <- subset(sub_Staten, room.and.type == "Private room")

#shared rooms
sub_shared_Man <- subset(sub_Manhattan, room.and.type == "Shared room")
sub_shared_Bronx <- subset(sub_Bronx, room.and.type == "Shared room")
sub_shared_Brooklyn <- subset(sub_Brooklyn, room.and.type == "Shared room")
sub_shared_Queens <- subset(sub_Queens, room.and.type == "Shared room")
sub_shared_Staten <- subset(sub_Staten, room.and.type == "Shared room")

#entire homes/apts
sub_entire_Man <- subset(sub_Manhattan, room.and.type == "Entire home/apt")
sub_entire_Bronx <- subset(sub_Bronx, room.and.type == "Entire home/apt")
sub_entire_Brooklyn <- subset(sub_Brooklyn, room.and.type == "Entire home/apt")
sub_entire_Queens <- subset(sub_Queens, room.and.type == "Entire home/apt")
sub_entire_Staten <- subset(sub_Staten, room.and.type == "Entire home/apt")
```

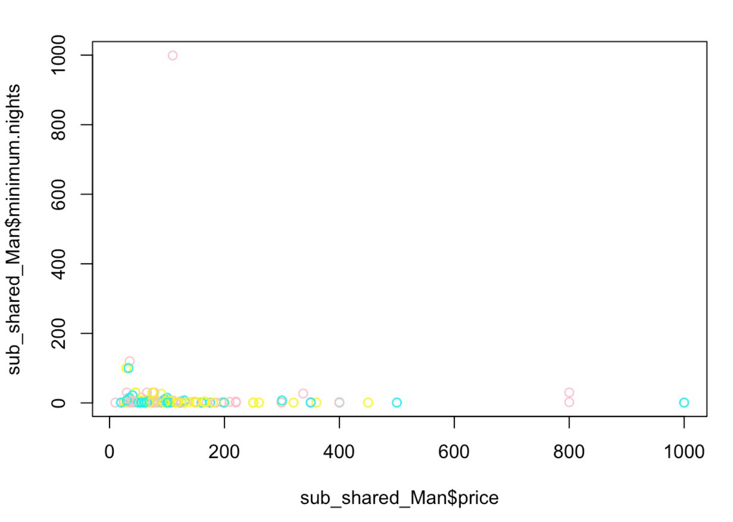
PRICE VS MIN NIGHTS IN MANHATTAN





While the observations remain same while looking specifically at private rooms in Manhattan, we observe many outliers here.

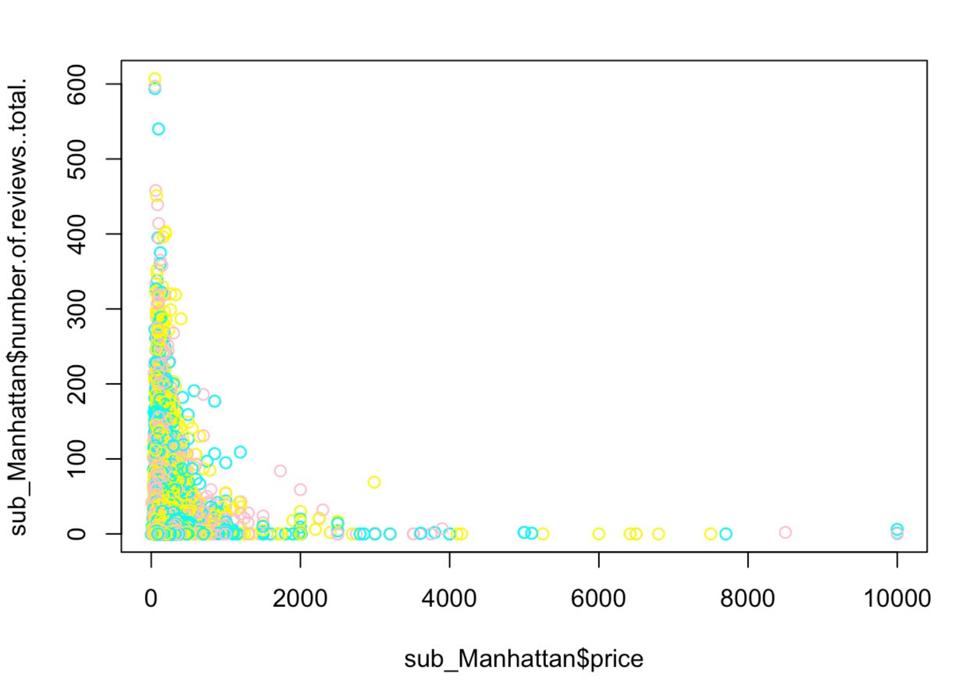
Places that cost less usually have lower number of minimum nights.



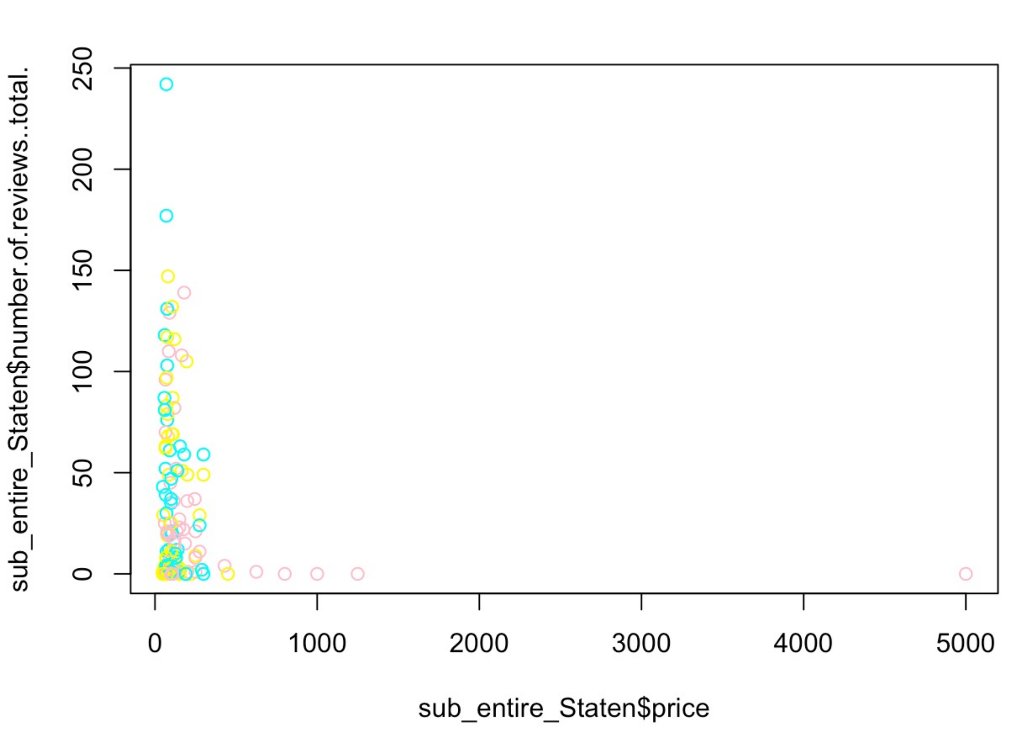
However, when we specifically look at all the shared spaces in Manhattan, we can see that regardless of price, the minimum nights remain the same.

This is a general trend with the relationship between price and minimum nights of a place, for other neighbourhood groups as well.

PRICE VS TOTAL NUM OF REVIEWS - MANHATTAN



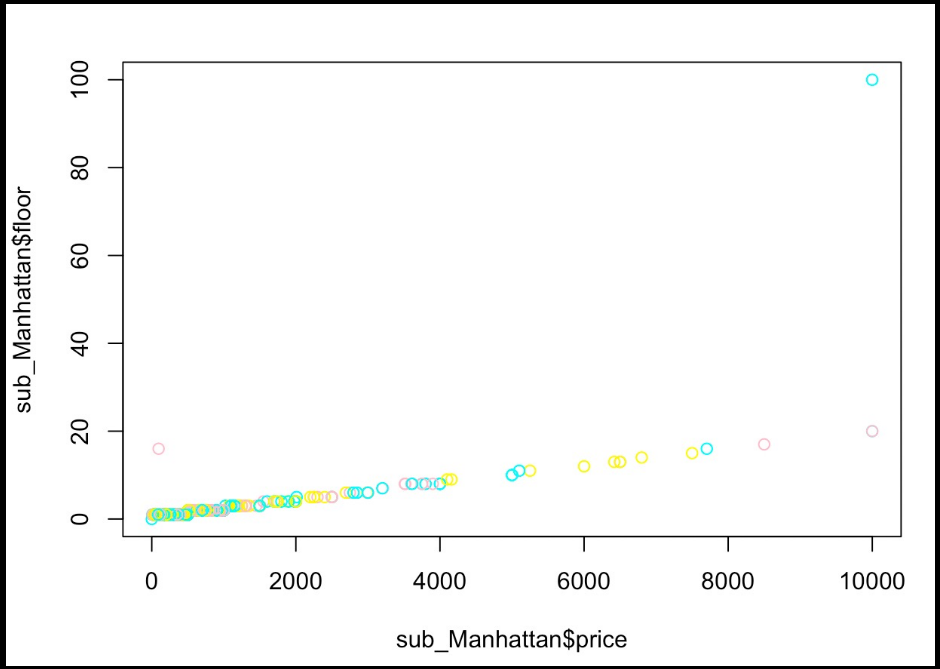
Places that cost less usually have a higher number of total reviews. This might be a result of the fact that since these places are more affordable, more people are able to get them, leading to a greater number of reviews.



Similar is the case for entire houses in Staten Island.

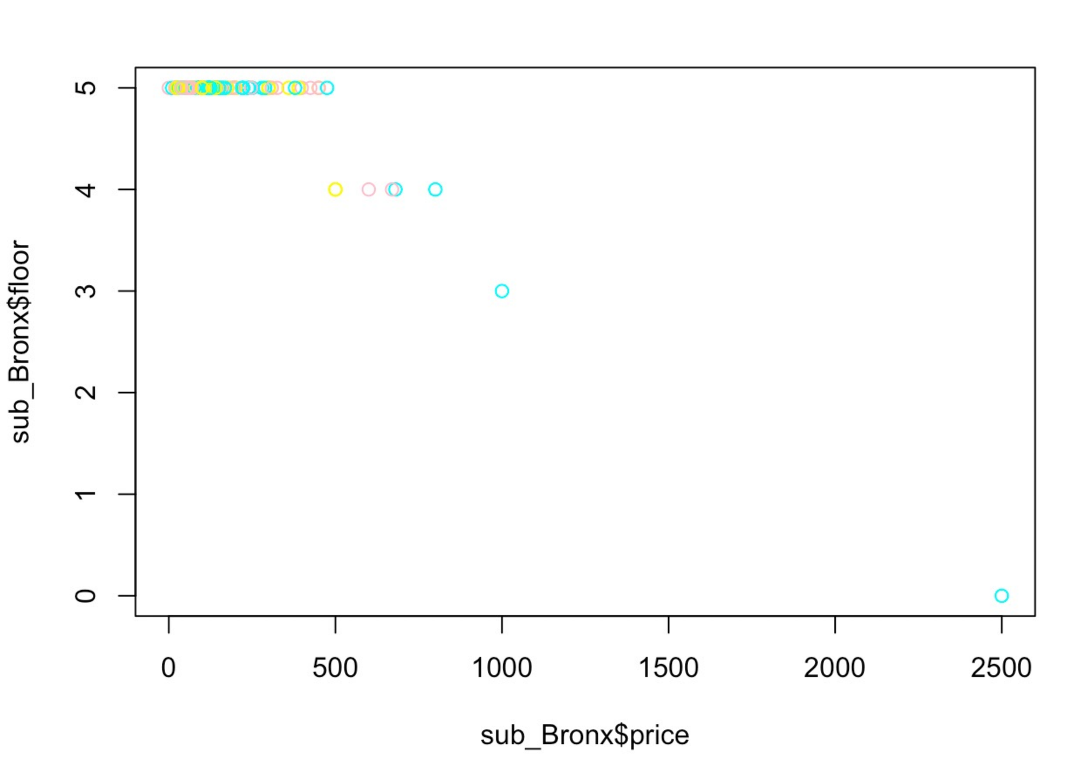
This is a general trend with the relationship between price and total number of reviews of a place, for other neighbourhood groups as well.

PRICE VS FLOOR - MANHATTAN

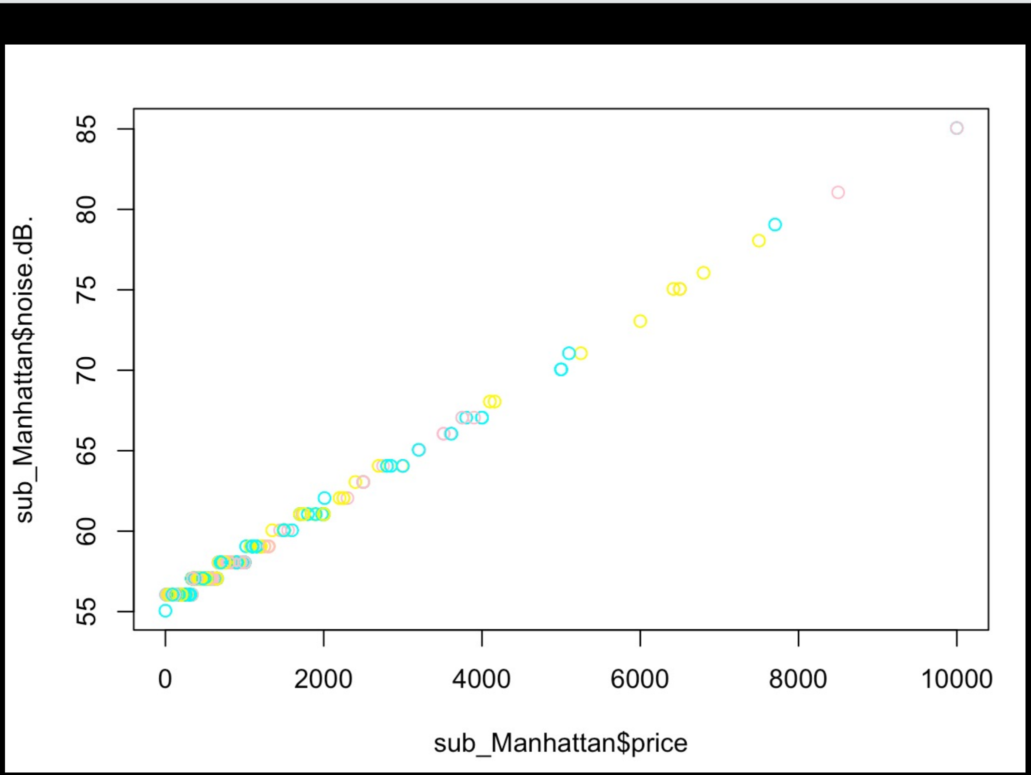


Here, we see a linear relationship between the price and the floors, indicating that the price increases as the level of floor increases – the higher the floor, the higher is the price.

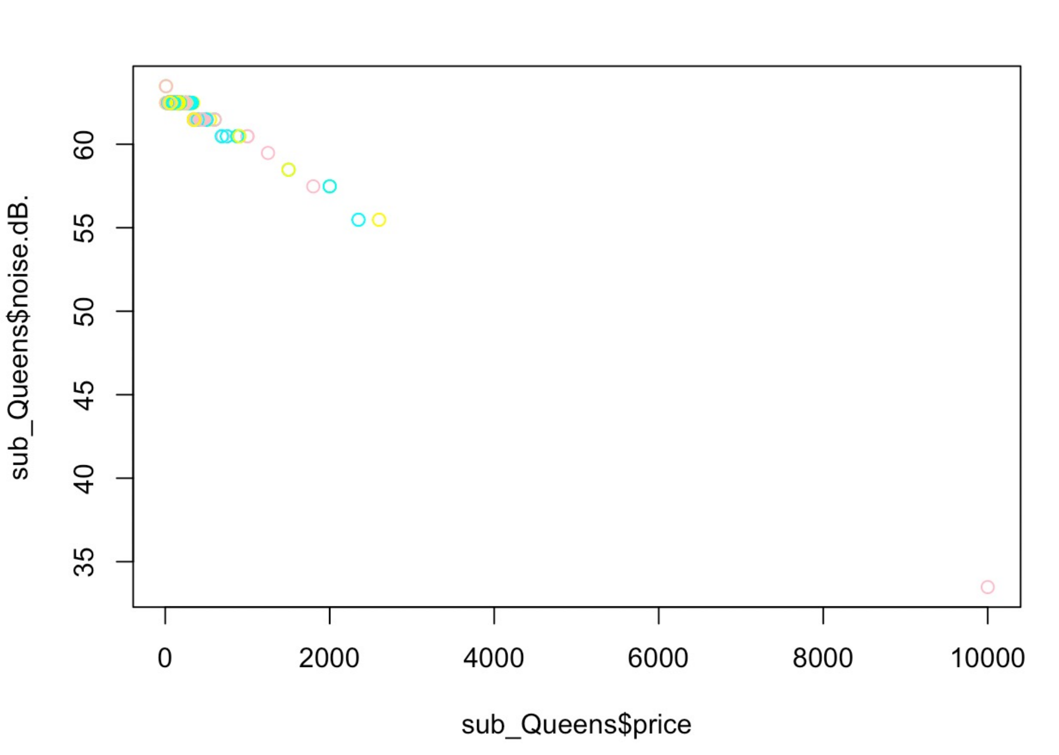
This however is not the case with Bronx. As we can see, most of the Airbnbs in Bronx are on the fifth floor, with a few outliers, thus giving us a very vague relationship bw floor and prices in Bronx.



PRICE VS NOISE - MANHATTAN



Such is not the case in Queens, because prices increases as noise levels decrease.



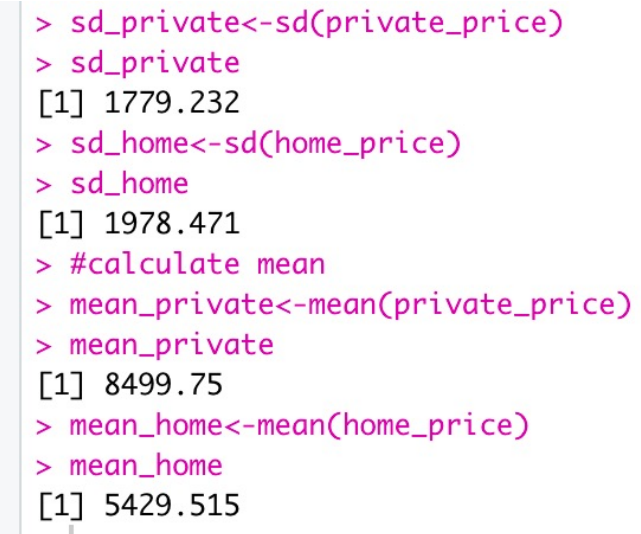
Here, we again see a linear relationship between the price and the noise levels, indicating that the price increases as the noise levels do, which although odd, makes sense in a city like Manhattan where everything is very lively.

HYPOTHESIS 1

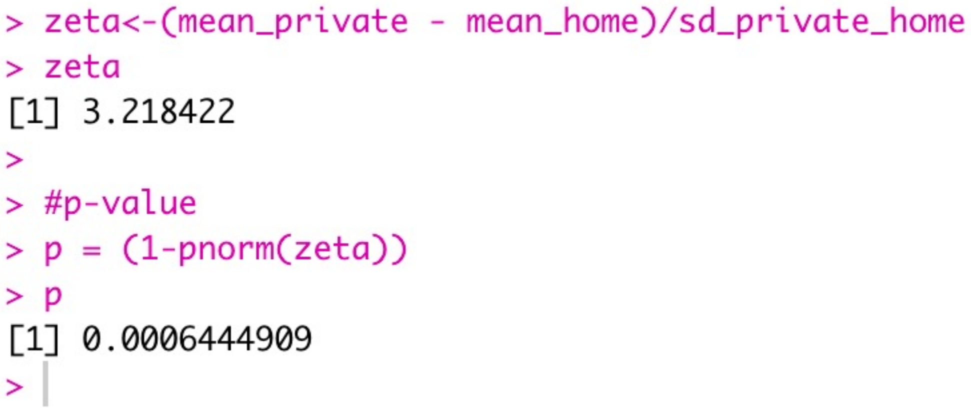
To test the previously mentioned hypothesis, I am going to find the p-value using z-test.

NULL HYPO: For high-end rooms, there is no difference in the affordability/prices of private rooms and entire homes/apts.

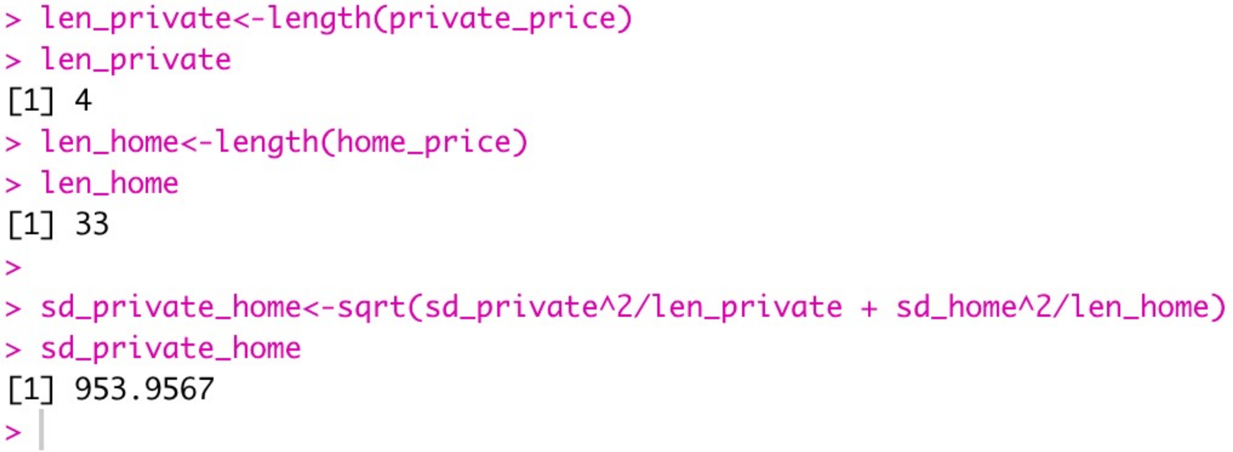
ALTERNATIVE HYPO: For high-end rooms, private rooms tend to be more expensive than entire homes/apts.



```
> sd_private<-sd(private_price)
> sd_private
[1] 1779.232
> sd_home<-sd(home_price)
> sd_home
[1] 1978.471
> #calculate mean
> mean_private<-mean(private_price)
> mean_private
[1] 8499.75
> mean_home<-mean(home_price)
> mean_home
[1] 5429.515
```



```
> zeta<-(mean_private - mean_home)/sd_private_home
> zeta
[1] 3.218422
>
> #p-value
> p = (1-pnorm(zeta))
> p
[1] 0.0006444909
>
```



```
> len_private<-length(private_price)
> len_private
[1] 4
> len_home<-length(home_price)
> len_home
[1] 33
>
> sd_private_home<-sqrt(sd_private^2/len_private + sd_home^2/len_home)
> sd_private_home
[1] 953.9567
>
```

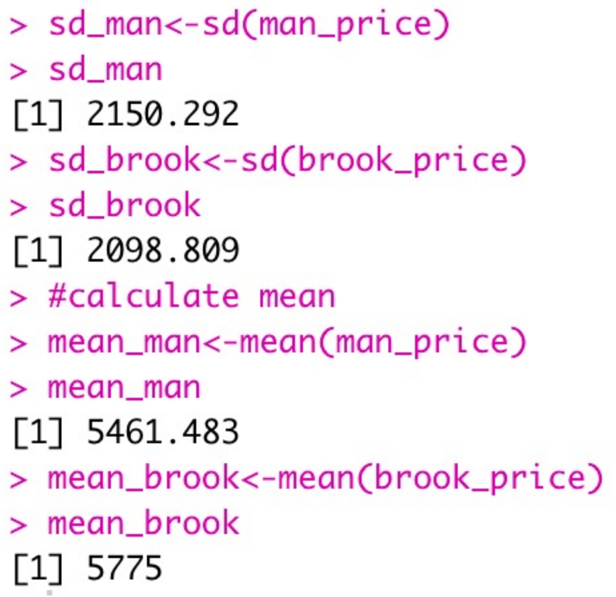
Following the z-test, since the p-value is less than 0.05, we REJECT the null hypothesis.

HYPOTHESIS 2

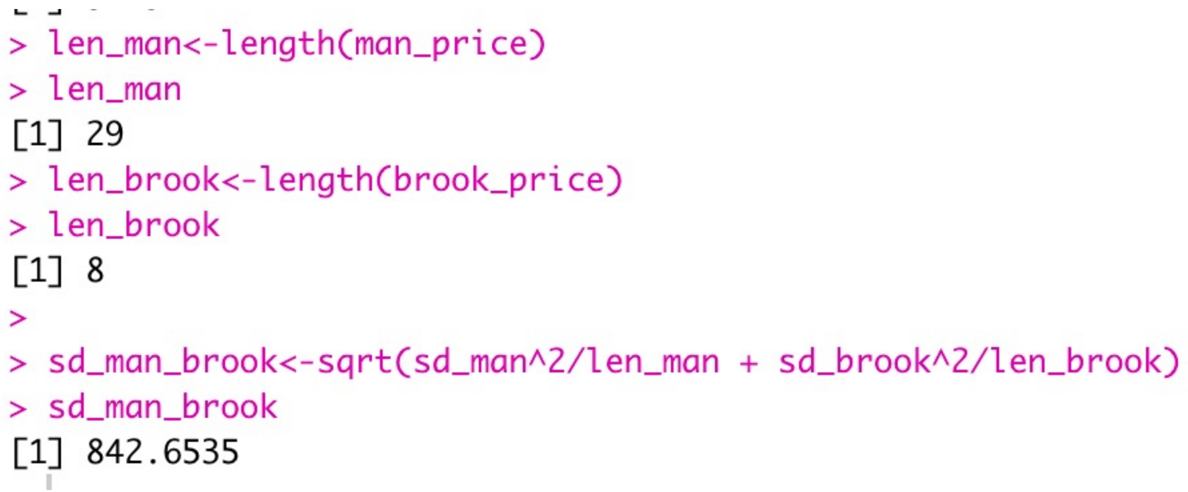
Looking at the previous plot, I wanted to test if Manhattan is more expensive than Brooklyn. I am going to find the p-value using z-test.

NULL HYPO: For high-end rooms, there is no difference in the affordability/prices of rooms in Manhattan and Brooklyn.

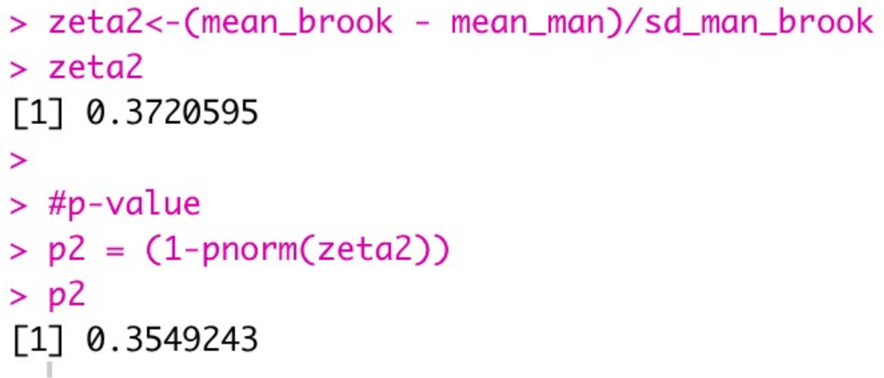
ALTERNATIVE HYPO: For high-end rooms, Manhattan has more expensive rooms than Brooklyn.



```
> sd_man<-sd(man_price)
> sd_man
[1] 2150.292
> sd_brook<-sd(brook_price)
> sd_brook
[1] 2098.809
> #calculate mean
> mean_man<-mean(man_price)
> mean_man
[1] 5461.483
> mean_brook<-mean(brook_price)
> mean_brook
[1] 5775
```



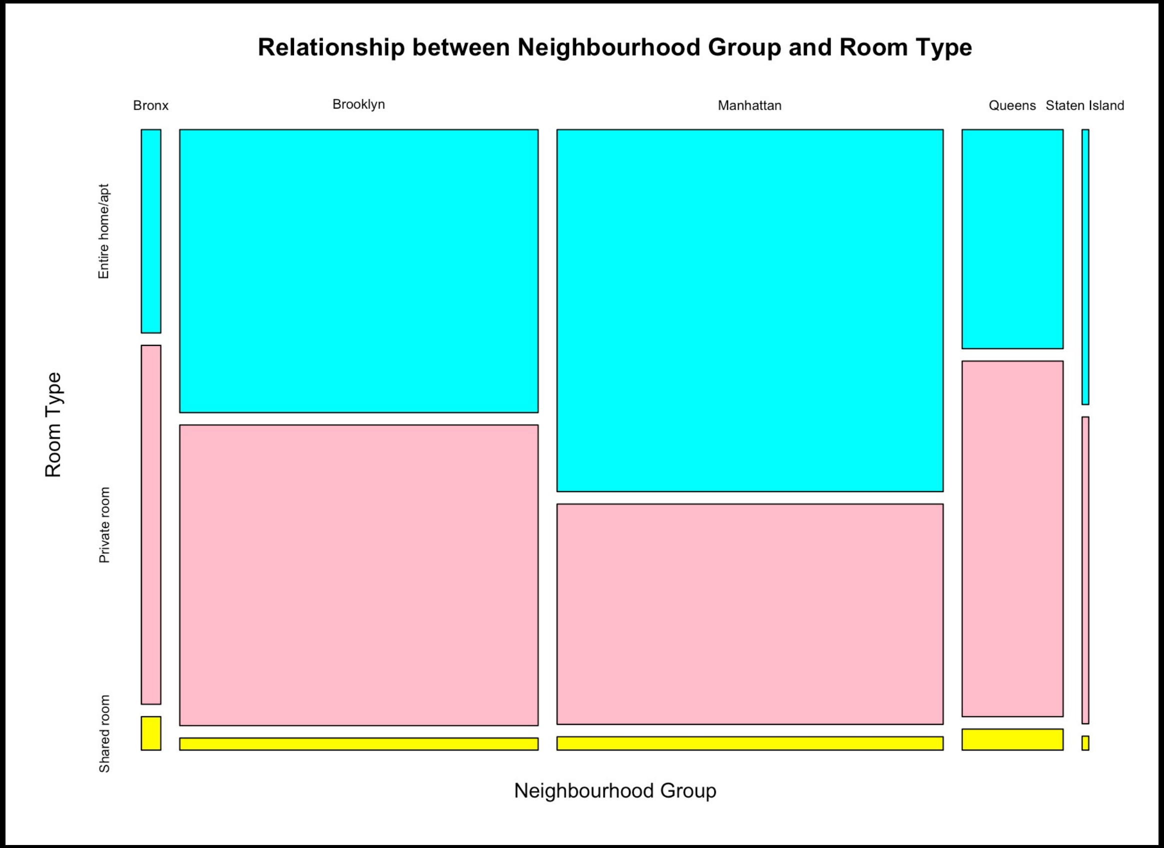
```
> len_man<-length(man_price)
> len_man
[1] 29
> len_brook<-length(brook_price)
> len_brook
[1] 8
>
> sd_man_brook<-sqrt(sd_man^2/len_man + sd_brook^2/len_brook)
> sd_man_brook
[1] 842.6535
```



```
> zeta2<-(mean_brook - mean_man)/sd_man_brook
> zeta2
[1] 0.3720595
>
> #p-value
> p2 = (1-pnorm(zeta2))
> p2
[1] 0.3549243
```

Following the z-test, since the p-value is greater than 0.05, we FAIL TO REJECT the null hypothesis.

RELATIONSHIP BW NEIGHBOURHOOD GROUP AND ROOM TYPE



Here is a quick look at the relationship between Neighbourhood Group and Room Type.

THANK YOU!