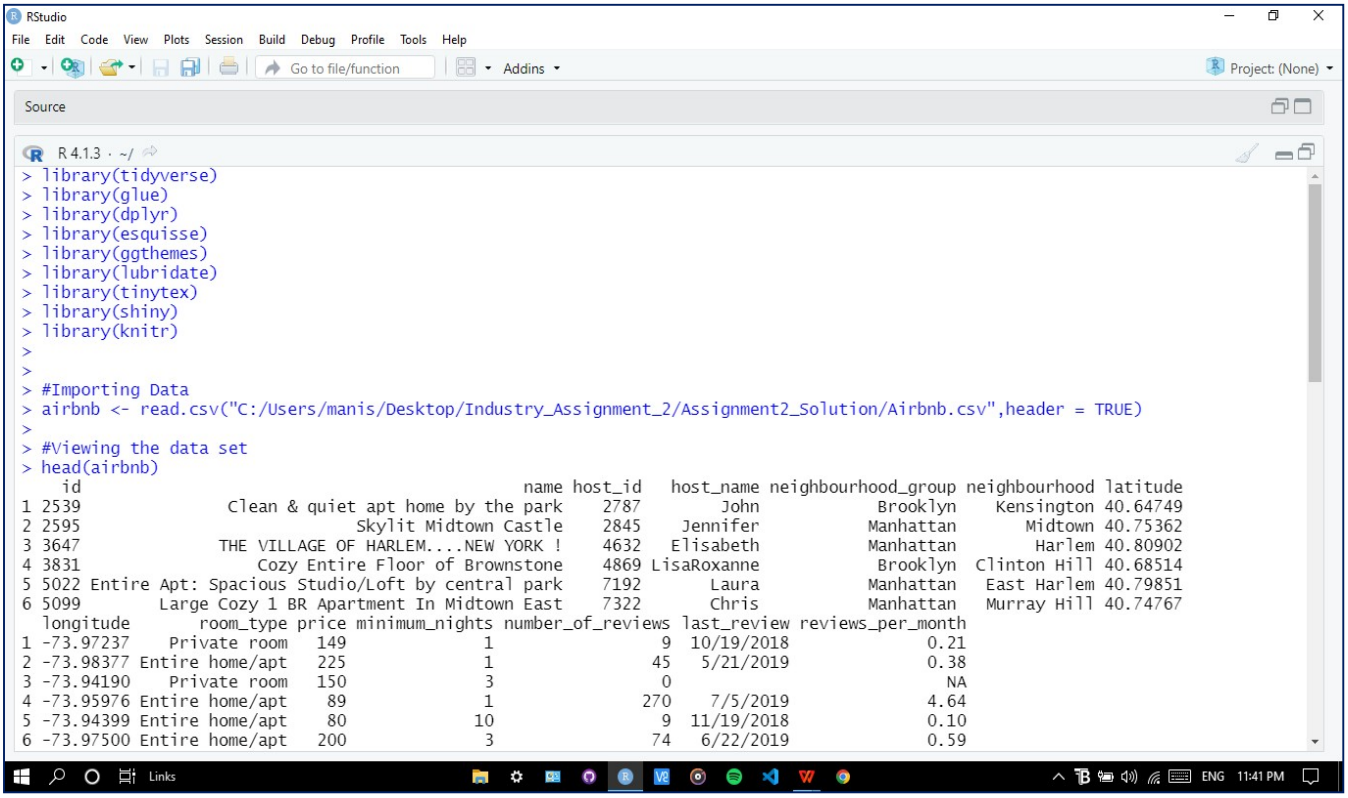
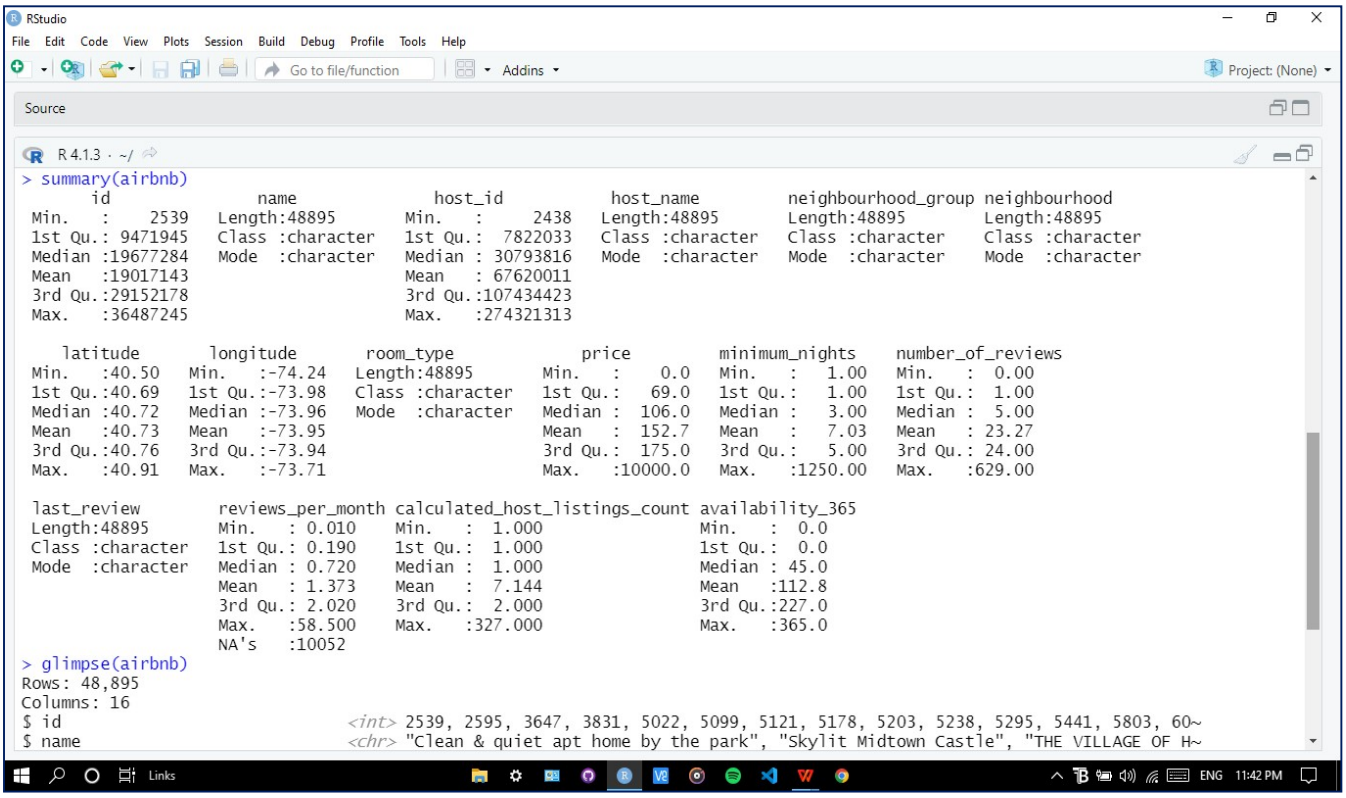
➢ Importing Some Required Library and the Data Set :



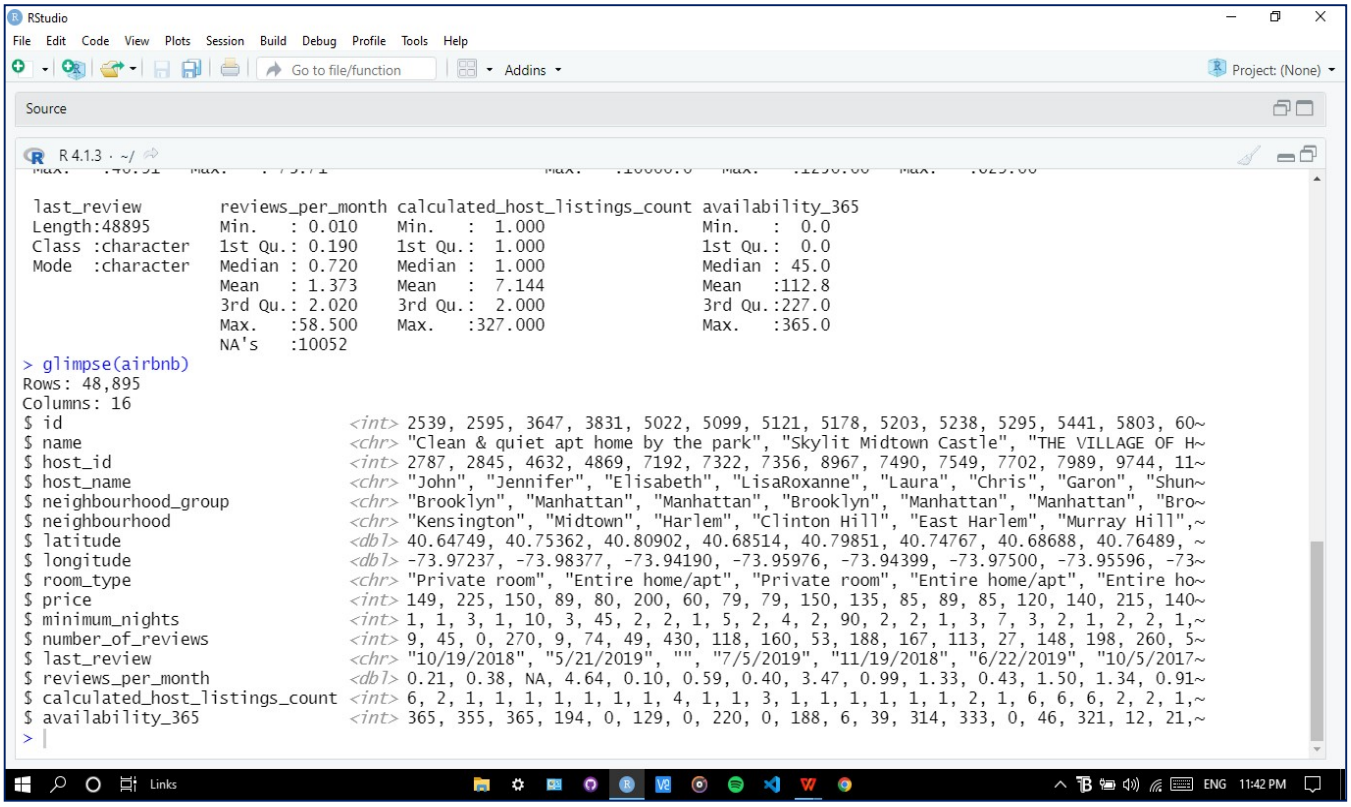
```
> library(tidyverse)
> library(glue)
> library(dplyr)
> library(esquisse)
> library(ggthemes)
> library(lubridate)
> library(tinytex)
> library(shiny)
> library(knitr)
>
>
> #Importing Data
> airbnb <- read.csv("C:/Users/manis/Desktop/Industry_Assignment_2/Assignment2_Solution/Airbnb.csv",header = TRUE)
>
> #Viewing the data set
> head(airbnb)
     id                                              name host_id   host_name neighbourhood_group neighbourhood latitude
1  2539                Clean & quiet apt home by the park    2787        John            Brooklyn    Kensington 40.64749
2  2595                             Skylit Midtown Castle    2845    Jennifer           Manhattan       Midtown 40.75362
3  3647               THE VILLAGE OF HARLEM....NEW YORK !    4632   Elisabeth           Manhattan        Harlem 40.80902
4  3831                   Cozy Entire Floor of Brownstone    4869 LisaRoxanne            Brooklyn  Clinton Hill 40.68514
5  5022 Entire Apt: Spacious Studio/Loft by central park    7192       Laura           Manhattan   East Harlem 40.79851
6  5099        Large Cozy 1 BR Apartment In Midtown East    7322       Chris           Manhattan   Murray Hill 40.74767
  longitude       room_type price minimum_nights number_of_reviews last_review reviews_per_month
1 -73.97237    Private room   149             1                 9  10/19/2018              0.21
2 -73.98377 Entire home/apt   225             1                45   5/21/2019              0.38
3 -73.94190    Private room   150             3                 0                          NA
4 -73.95976 Entire home/apt    89             1               270   7/5/2019               4.64
5 -73.94399 Entire home/apt    80            10                 9  11/19/2018              0.10
6 -73.97500 Entire home/apt   200             3                74   6/22/2019              0.59
```



```
> summary(airbnb)
       id                 name              host_id           host_name         neighbourhood_group neighbourhood
 Min.   :    2539   Length:48895       Min.   :     2438   Length:48895       Length:48895        Length:48895
 1st Qu.: 9471945   Class :character   1st Qu.:  7822033   Class :character   Class :character    Class :character
 Median :19677284   Mode  :character   Median : 30793816   Mode  :character   Mode  :character    Mode  :character
 Mean   :19017143                      Mean   : 67620011
 3rd Qu.:29152178                      3rd Qu.:107434423
 Max.   :36487245                      Max.   :274321313

    latitude       longitude       room_type             price         minimum_nights    number_of_reviews
 Min.   :40.50   Min.   :-74.24   Length:48895       Min.   :    0.0   Min.   :   1.00   Min.   :  0.00
 1st Qu.:40.69   1st Qu.:-73.98   Class :character   1st Qu.:   69.0   1st Qu.:   1.00   1st Qu.:  1.00
 Median :40.72   Median :-73.96   Mode  :character   Median :  106.0   Median :   3.00   Median :  5.00
 Mean   :40.73   Mean   :-73.95                      Mean   :  152.7   Mean   :   7.03   Mean   : 23.27
 3rd Qu.:40.76   3rd Qu.:-73.94                      3rd Qu.:  175.0   3rd Qu.:   5.00   3rd Qu.: 24.00
 Max.   :40.91   Max.   :-73.71                      Max.   :10000.0   Max.   :1250.00   Max.   :629.00

 last_review        reviews_per_month calculated_host_listings_count availability_365
 Length:48895       Min.   : 0.010    Min.   :  1.000               Min.   :  0.0
 Class :character   1st Qu.: 0.190    1st Qu.:  1.000               1st Qu.:  0.0
 Mode  :character   Median : 0.720    Median :  1.000               Median : 45.0
                    Mean   : 1.373    Mean   :  7.144               Mean   :112.8
                    3rd Qu.: 2.020    3rd Qu.:  2.000               3rd Qu.:227.0
                    Max.   :58.500    Max.   :327.000               Max.   :365.0
                    NA's   :10052
> glimpse(airbnb)
Rows: 48,895
Columns: 16
$ id                            <int> 2539, 2595, 3647, 3831, 5022, 5099, 5121, 5178, 5203, 5238, 5295, 5441, 5803, 60~
$ name                          <chr> "Clean & quiet apt home by the park", "Skylit Midtown Castle", "THE VILLAGE OF H~
```

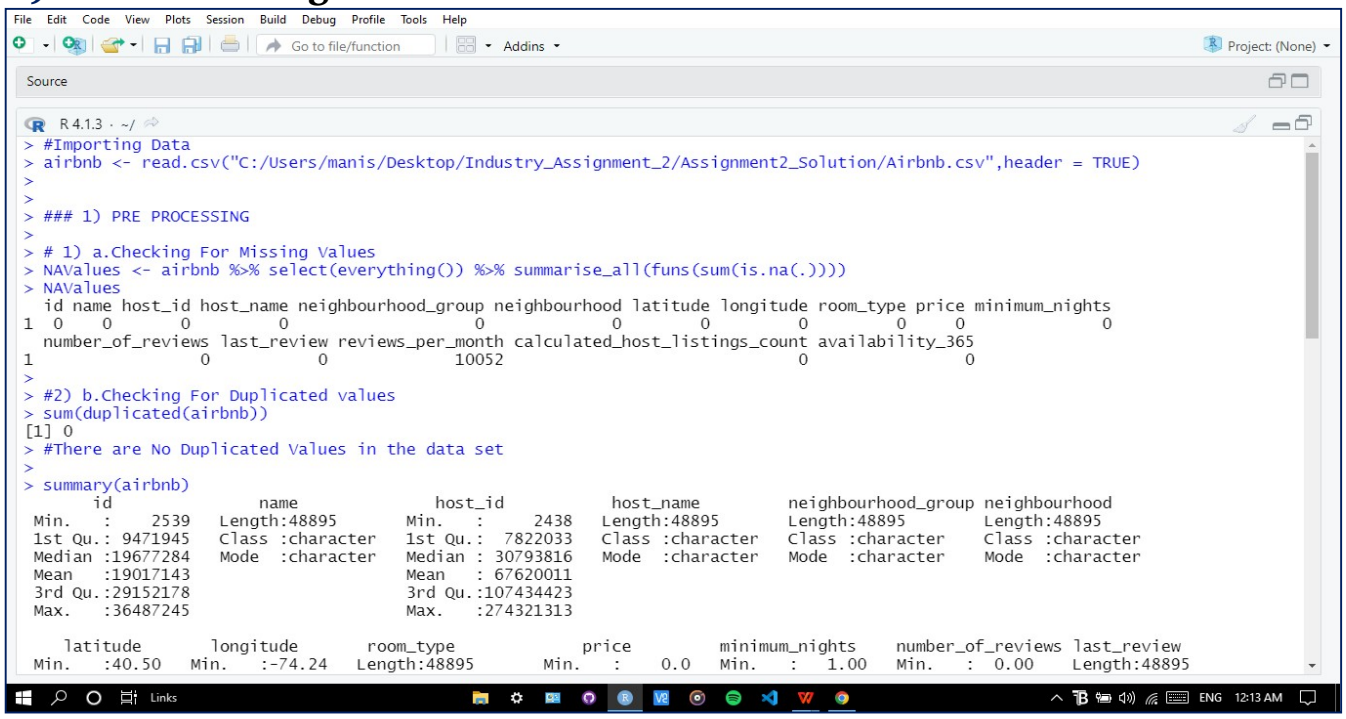


- **Analyzing the Data Set to Perform :**

1) **Pre - Processing**
2) **Exploratory Data Analysis.**
3) **Plot Graphs Based On SomeConditions.**

1) Pre - Processing :

a) Treat missing values :

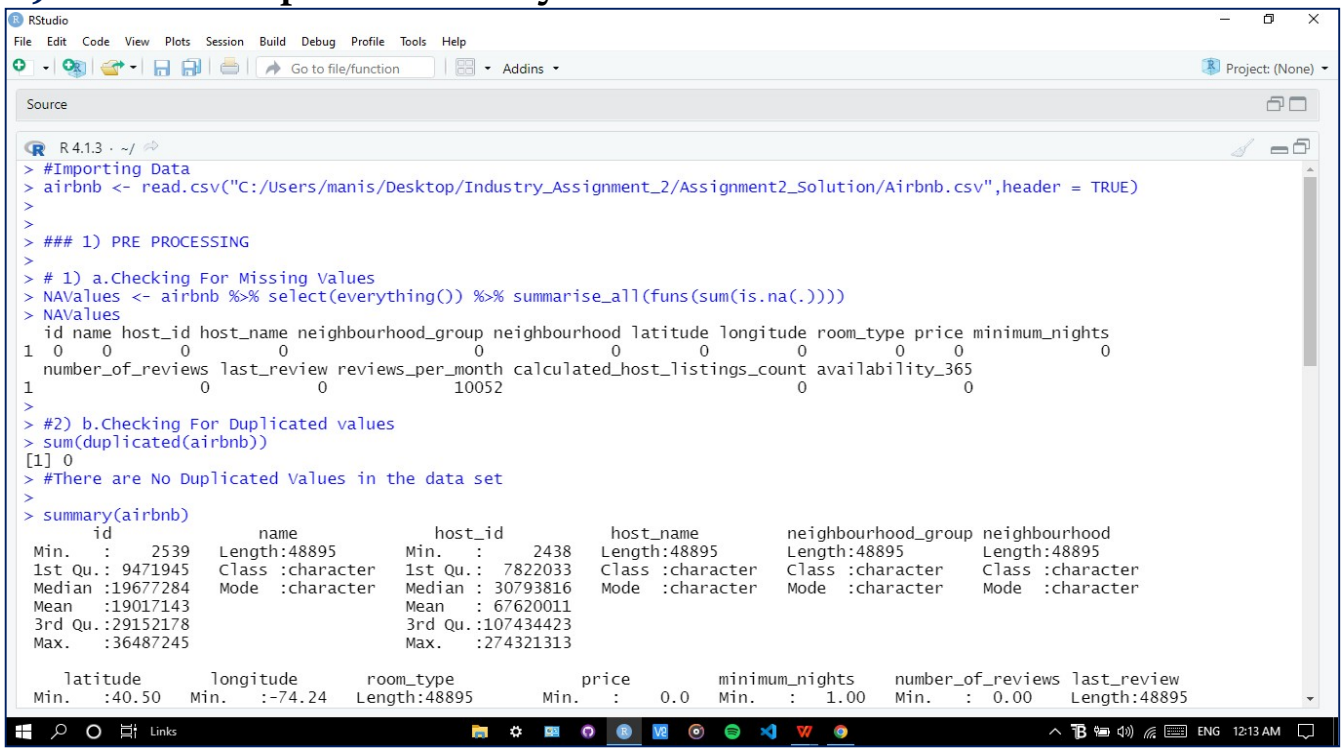


```
> #Importing Data
> airbnb <- read.csv("C:/Users/manis/Desktop/Industry_Assignment_2/Assignment2_Solution/Airbnb.csv",header = TRUE)
>
>
> ### 1) PRE PROCESSING
>
> # 1) a.Checking For Missing Values
> NAValues <- airbnb %>% select(everything()) %>% summarise_all(funs(sum(is.na(.))))
> NAValues
  id name host_id host_name neighbourhood_group neighbourhood latitude longitude room_type price minimum_nights
1  0    0       0         0                   0             0        0         0         0     0              0
  number_of_reviews last_review reviews_per_month calculated_host_listings_count availability_365
1                 0           0            10052                               0               0
>
> #2) b.Checking For Duplicated values
> sum(duplicated(airbnb))
[1] 0
> #There are No Duplicated Values in the data set
>
> summary(airbnb)
       id                name              host_id           host_name         neighbourhood_group neighbourhood
 Min.   :    2539   Length:48895       Min.   :     2438   Length:48895       Length:48895        Length:48895
 1st Qu.: 9471945   Class :character   1st Qu.:  7822033   Class :character   Class :character    Class :character
 Median :19677284   Mode  :character   Median : 30793816   Mode  :character   Mode  :character    Mode  :character
 Mean   :19017143                      Mean   : 67620011
 3rd Qu.:29152178                      3rd Qu.:107434423
 Max.   :36487245                      Max.   :274321313

    latitude       longitude       room_type             price         minimum_nights    number_of_reviews last_review
 Min.   :40.50   Min.   :-74.24   Length:48895       Min.   :    0.0   Min.   :   1.00   Min.   :  0.00    Length:48895
```

#There are Total 10052 values missing from reviews per month column

b) Remove duplicates if any :



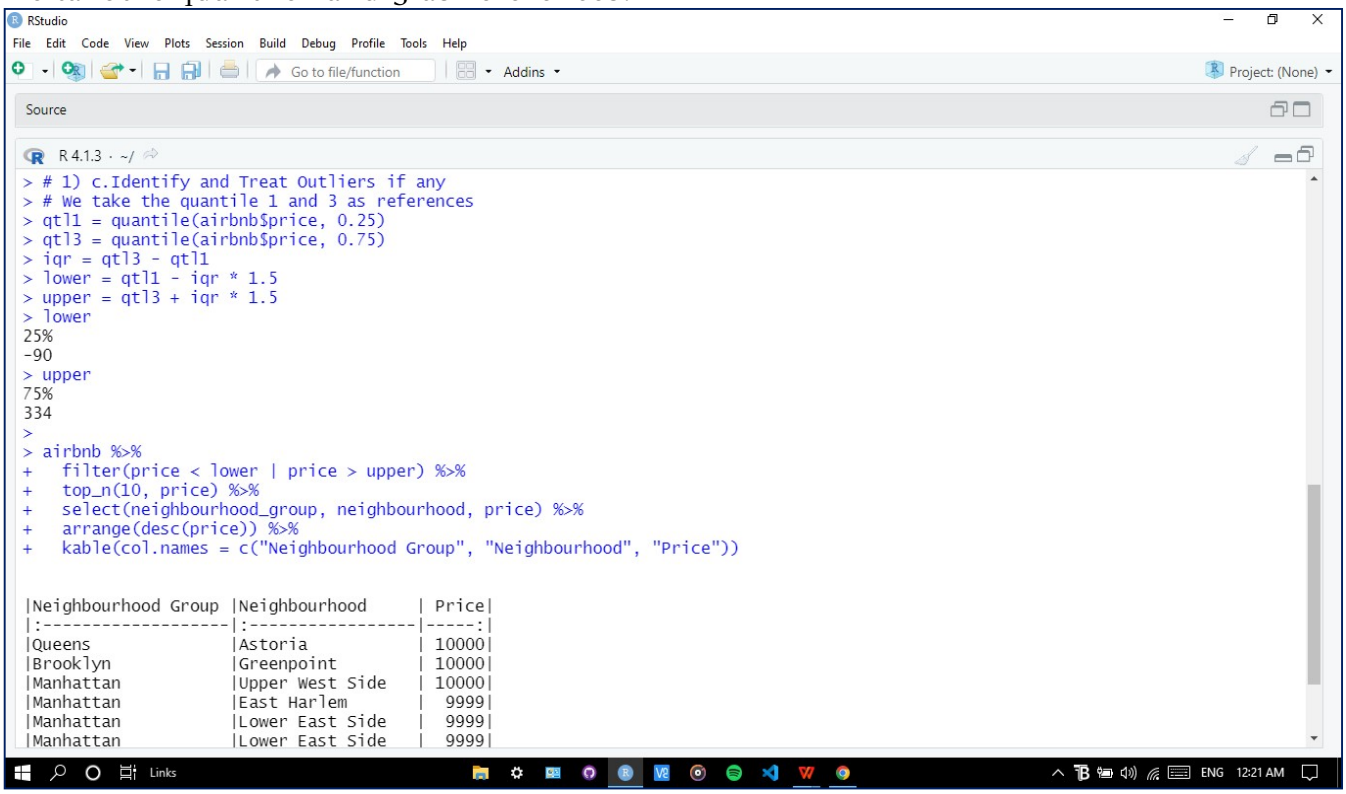
```
> #Importing Data
> airbnb <- read.csv("C:/Users/manis/Desktop/Industry_Assignment_2/Assignment2_Solution/Airbnb.csv",header = TRUE)
>
>
> ### 1) PRE PROCESSING
>
> # 1) a.Checking For Missing Values
> NAValues <- airbnb %>% select(everything()) %>% summarise_all(funs(sum(is.na(.))))
> NAValues
  id name host_id host_name neighbourhood_group neighbourhood latitude longitude room_type price minimum_nights
1  0    0       0         0                   0             0        0         0         0     0              0
  number_of_reviews last_review reviews_per_month calculated_host_listings_count availability_365
1                 0           0            10052                               0               0
>
> #2) b.Checking For Duplicated values
> sum(duplicated(airbnb))
[1] 0
> #There are No Duplicated Values in the data set
>
> summary(airbnb)
       id                name              host_id           host_name         neighbourhood_group neighbourhood
 Min.   :    2539   Length:48895       Min.   :     2438   Length:48895       Length:48895        Length:48895
 1st Qu.: 9471945   Class :character   1st Qu.:  7822033   Class :character   Class :character    Class :character
 Median :19677284   Mode  :character   Median : 30793816   Mode  :character   Mode  :character    Mode  :character
 Mean   :19017143                      Mean   : 67620011
 3rd Qu.:29152178                      3rd Qu.:107434423
 Max.   :36487245                      Max.   :274321313

    latitude       longitude       room_type             price         minimum_nights    number_of_reviews last_review
 Min.   :40.50   Min.   :-74.24   Length:48895       Min.   :    0.0   Min.   :   1.00   Min.   :  0.00    Length:48895
```

#There are No Duplicated Values in the data set

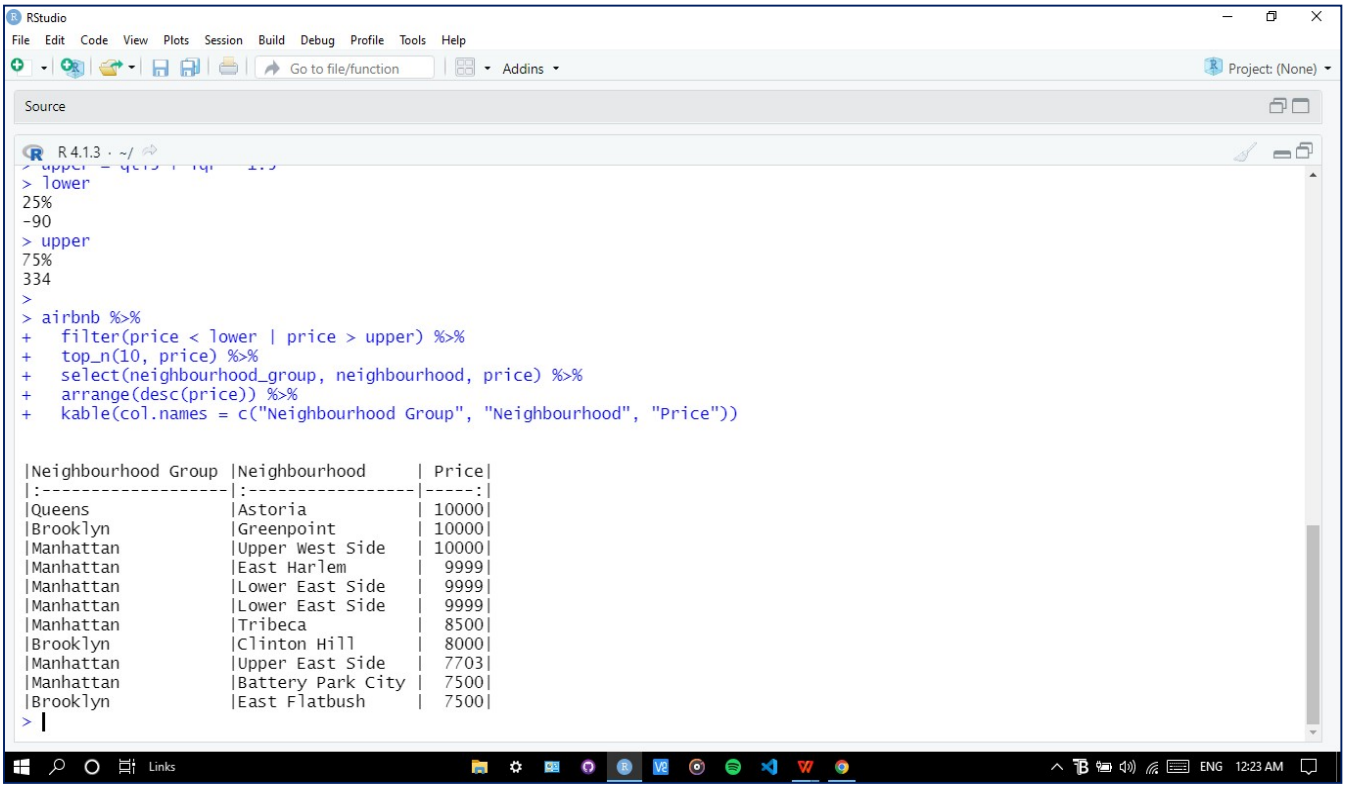
c) Identify and treat outliers if any :

Before starting our analysis, we also want to check the outlier points in this data set and we takethe quantile1 and 3 as references.



```
> # 1) c.Identify and Treat Outliers if any
> # We take the quantile 1 and 3 as references
> qtl1 = quantile(airbnb$price, 0.25)
> qtl3 = quantile(airbnb$price, 0.75)
> iqr = qtl3 - qtl1
> lower = qtl1 - iqr * 1.5
> upper = qtl3 + iqr * 1.5
> lower
25%
-90
> upper
75%
334
>
> airbnb %>%
+   filter(price < lower | price > upper) %>%
+   top_n(10, price) %>%
+   select(neighbourhood_group, neighbourhood, price) %>%
+   arrange(desc(price)) %>%
+   kable(col.names = c("Neighbourhood Group", "Neighbourhood", "Price"))


|Neighbourhood Group |Neighbourhood     | Price|
|:-------------------|:-----------------|-----:|
|Queens              |Astoria           | 10000|
|Brooklyn            |Greenpoint        | 10000|
|Manhattan           |Upper West Side   | 10000|
|Manhattan           |East Harlem       |  9999|
|Manhattan           |Lower East Side   |  9999|
|Manhattan           |Lower East Side   |  9999|
```



```
> lower
25%
-90
> upper
75%
334
>
> airbnb %>%
+   filter(price < lower | price > upper) %>%
+   top_n(10, price) %>%
+   select(neighbourhood_group, neighbourhood, price) %>%
+   arrange(desc(price)) %>%
+   kable(col.names = c("Neighbourhood Group", "Neighbourhood", "Price"))


|Neighbourhood Group |Neighbourhood     | Price|
|:-------------------|:-----------------|-----:|
|Queens              |Astoria           | 10000|
|Brooklyn            |Greenpoint        | 10000|
|Manhattan           |Upper West Side   | 10000|
|Manhattan           |East Harlem       |  9999|
|Manhattan           |Lower East Side   |  9999|
|Manhattan           |Lower East Side   |  9999|
|Manhattan           |Tribeca           |  8500|
|Brooklyn            |Clinton Hill      |  8000|
|Manhattan           |Upper East Side   |  7703|
|Manhattan           |Battery Park City |  7500|
|Brooklyn            |East Flatbush     |  7500|
>
```

d) Transform Data based on judgement and explain themethodology used :

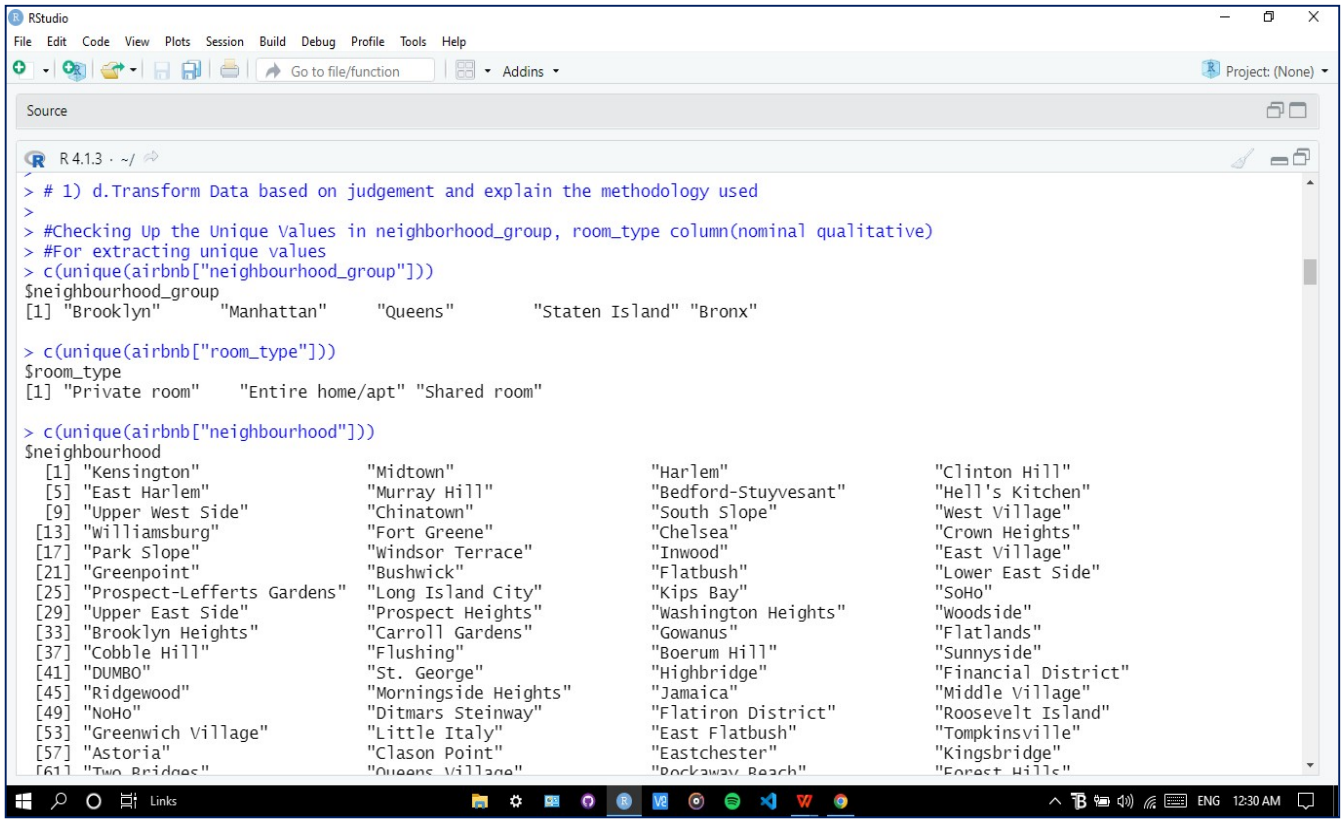
Methodology I Used :

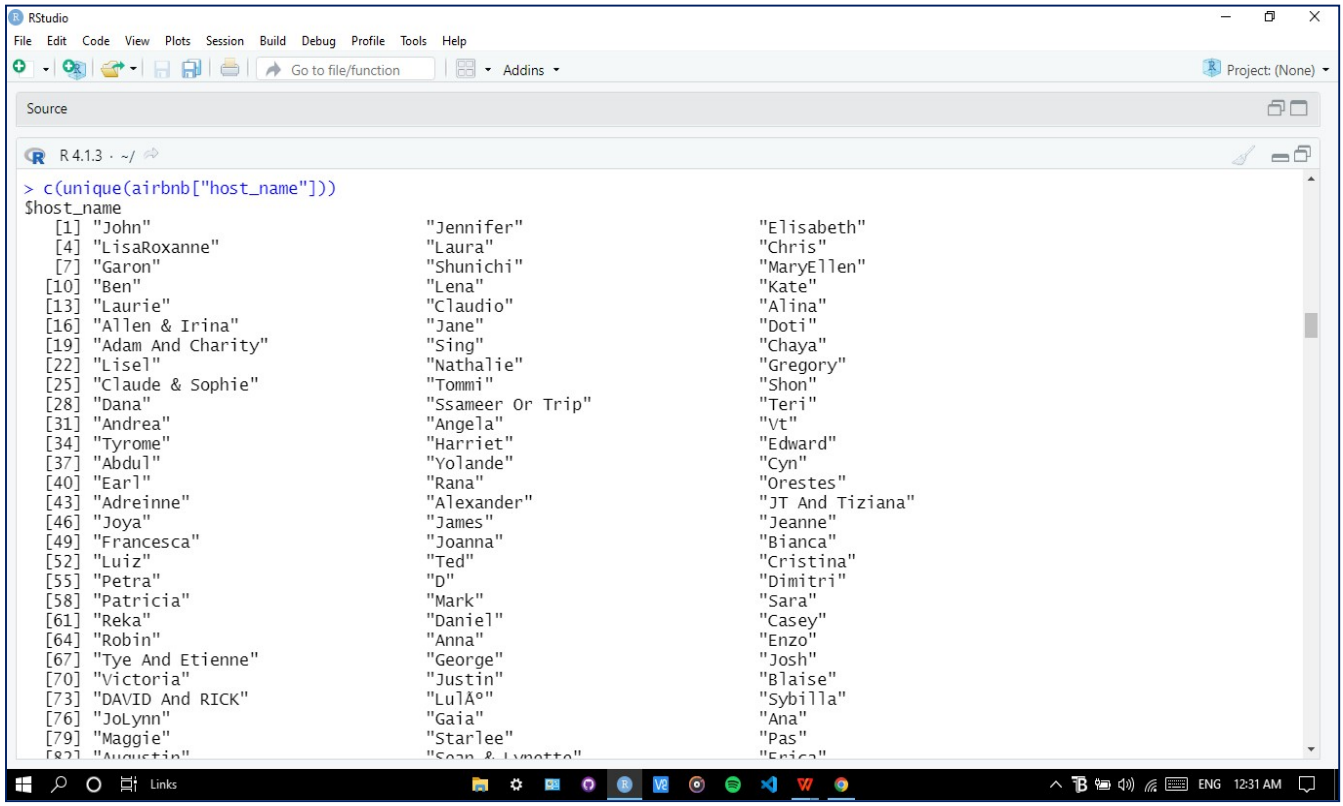
The methodology I have used here is that first I have checked the unique values in neighborhood_group, room_type, neighbourhood, host_name, and minimum_nights and then I have converted the neighbourhood_group, neighbourhood and room type to factors to see the types of it with the help of level() function.

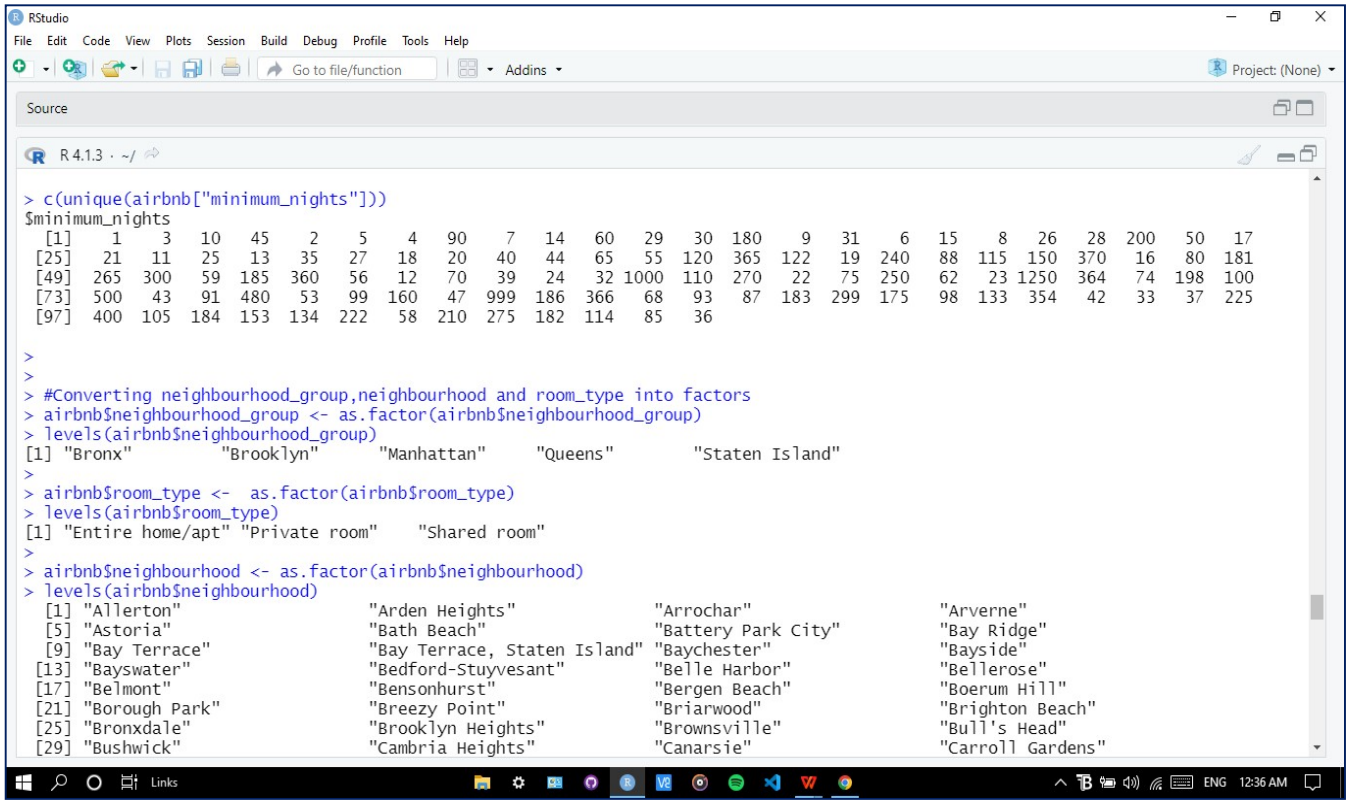
In the data set the last review column is not properly defined, so in order to remove the complexity there should be a specific format. I have separated the dates with slashes (/), saved them by Month,Day, and Year , and then replaced NA values with zero (o), then convertedthem back to integers!

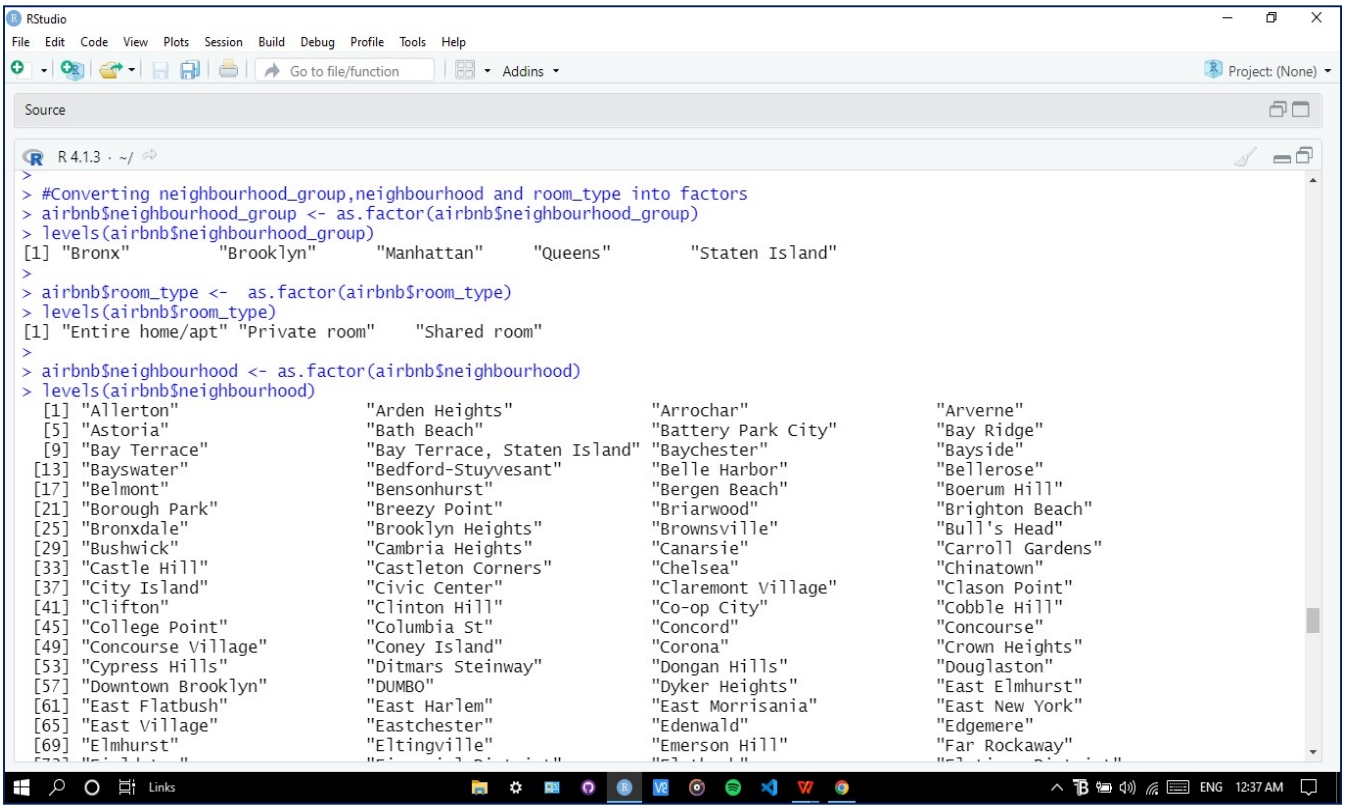
Replacing NA values in review_per_month column with zero(o).

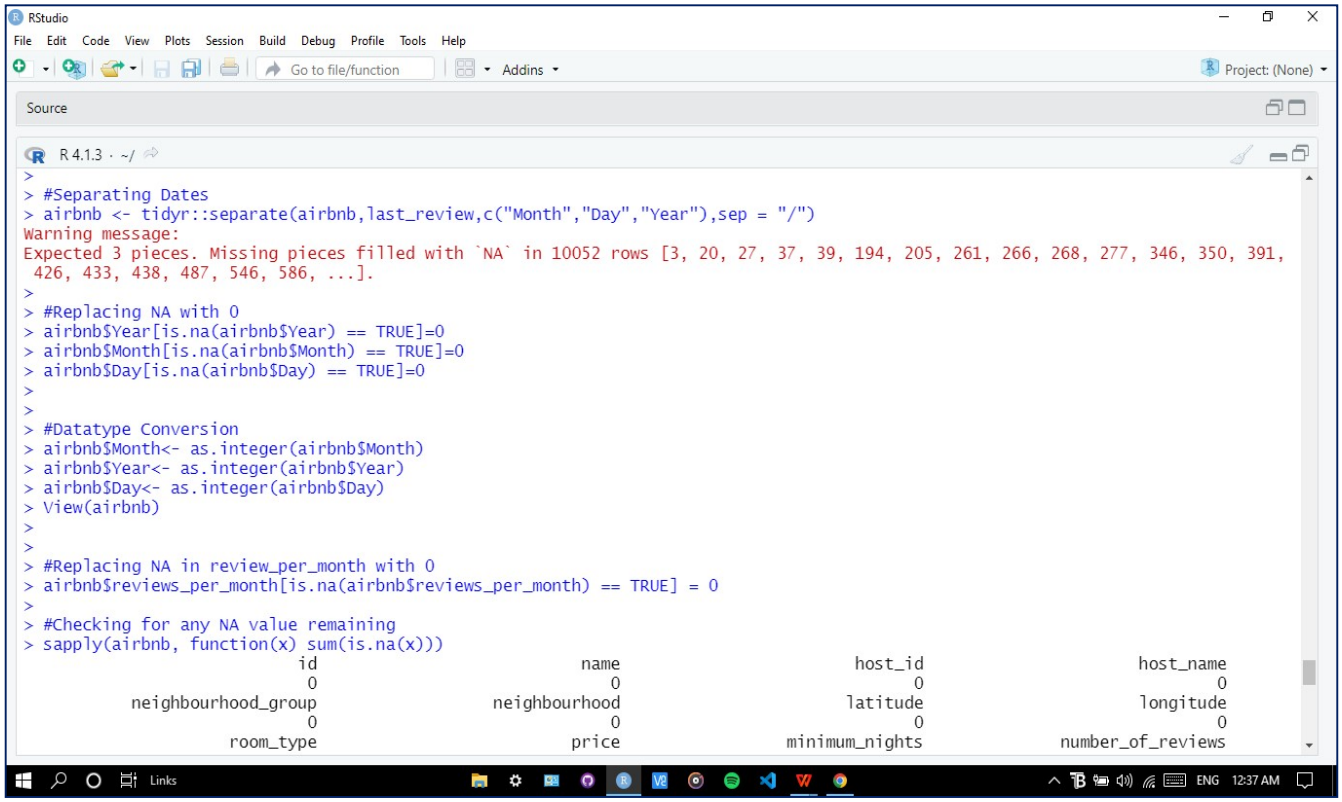
Viewing up the summary!


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Source
R 4.1.3 · ~/
> # 1) d.Transform Data based on judgement and explain the methodology used
>
> #Checking Up the Unique Values in neighborhood_group, room_type column(nominal qualitative)
> #For extracting unique values
> c(unique(airbnb["neighbourhood_group"]))
$neighbourhood_group
[1] "Brooklyn" "Manhattan" "Queens" "Staten Island" "Bronx"
> c(unique(airbnb["room_type"]))
$room_type
[1] "Private room" "Entire home/apt" "Shared room"
> c(unique(airbnb["neighbourhood"]))
$neighbourhood
[1] "Kensington" "Midtown" "Harlem" "Clinton Hill"
[5] "East Harlem" "Murray Hill" "Bedford-Stuyvesant" "Hell's Kitchen"
[9] "Upper West Side" "Chinatown" "South Slope" "West Village"
[13] "Williamsburg" "Fort Greene" "Chelsea" "Crown Heights"
[17] "Park Slope" "Windsor Terrace" "Inwood" "East Village"
[21] "Greenpoint" "Bushwick" "Flatbush" "Lower East Side"
[25] "Prospect-Lefferts Gardens" "Long Island City" "Kips Bay" "SoHo"
[29] "Upper East Side" "Prospect Heights" "Washington Heights" "Woodside"
[33] "Brooklyn Heights" "Carroll Gardens" "Gowanus" "Flatlands"
[37] "Cobble Hill" "Flushing" "Boerum Hill" "Sunnyside"
[41] "DUMBO" "St. George" "Highbridge" "Financial District"
[45] "Ridgewood" "Morningside Heights" "Jamaica" "Middle Village"
[49] "NoHo" "Ditmars Steinway" "Flatiron District" "Roosevelt Island"
[53] "Greenwich Village" "Little Italy" "East Flatbush" "Tompkinsville"
[57] "Astoria" "Clason Point" "Eastchester" "Kingsbridge"
Links
ENG 12:30 AM

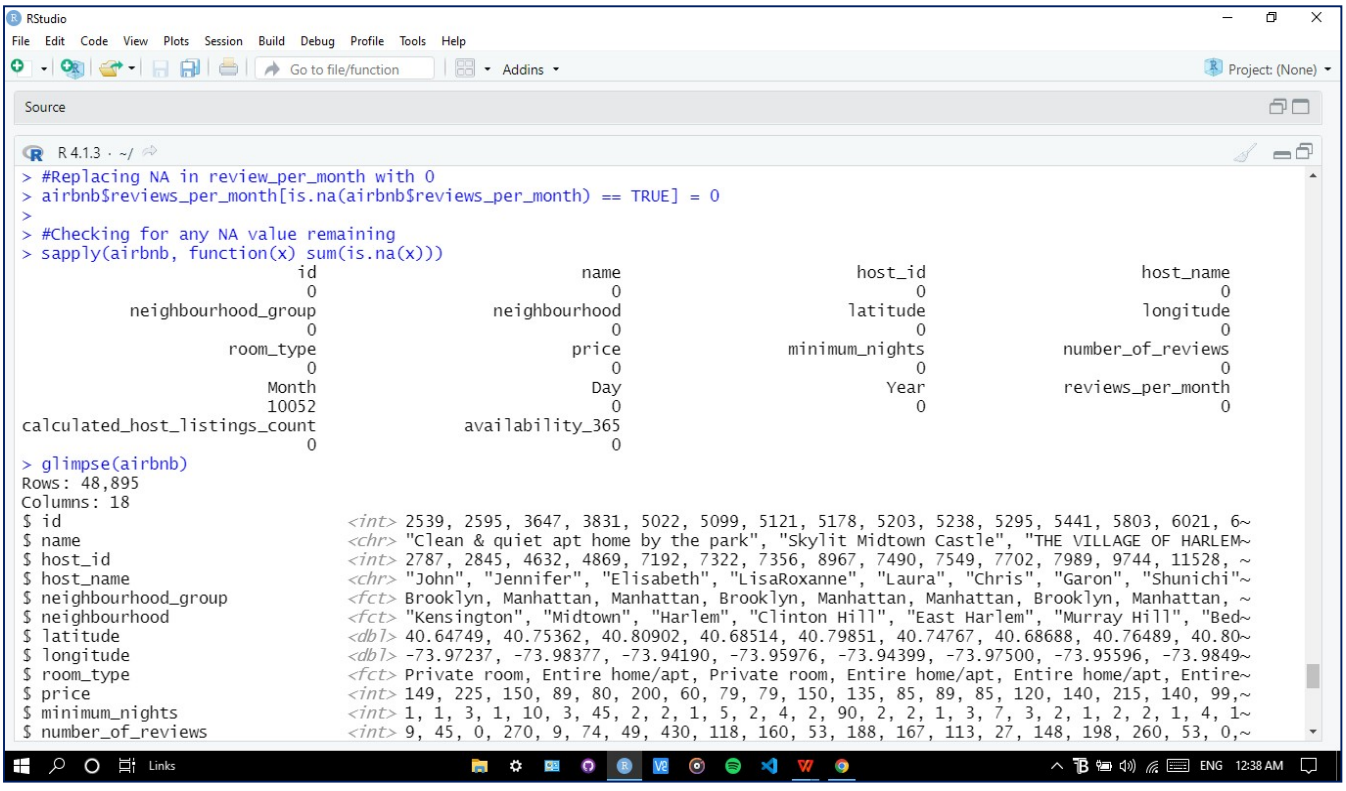

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Source
R 4.1.3 · ~/
> c(unique(airbnb["host_name"]))
$host_name
[1] "John" "Jennifer" "Elisabeth"
[4] "LisaRoxanne" "Laura" "Chris"
[7] "Garon" "Shunichi" "MaryEllen"
[10] "Ben" "Lena" "Kate"
[13] "Laurie" "Claudio" "Alina"
[16] "Allen & Irina" "Jane" "Doti"
[19] "Adam And Charity" "Sing" "Chaya"
[22] "Lisel" "Nathalie" "Gregory"
[25] "Claude & Sophie" "Tommi" "Shon"
[28] "Dana" "Ssameer Or Trip" "Teri"
[31] "Andrea" "Angela" "Vt"
[34] "Tyrome" "Harriet" "Edward"
[37] "Abdul" "Yolande" "Cyn"
[40] "Earl" "Rana" "Orestes"
[43] "Adreinne" "Alexander" "JT And Tiziana"
[46] "Joya" "James" "Jeanne"
[49] "Francesca" "Joanna" "Bianca"
[52] "Luiz" "Ted" "Cristina"
[55] "Petra" "D" "Dimitri"
[58] "Patricia" "Mark" "Sara"
[61] "Reka" "Daniel" "Casey"
[64] "Robin" "Anna" "Enzo"
[67] "Tye And Etienne" "George" "Josh"
[70] "Victoria" "Justin" "Blaise"
[73] "DAVID And RICK" "LulÃº" "Sybilla"
[76] "JoLynn" "Gaia" "Ana"
[79] "Maggie" "Starlee" "Pas"
Links
ENG 12:31 AM


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Source
R 4.1.3 · ~/
> c(unique(airbnb["minimum_nights"]))
$minimum_nights
[1] 1 3 10 45 2 5 4 90 7 14 60 29 30 180 9 31 6 15 8 26 28 200 50 17
[25] 21 11 25 13 35 27 18 20 40 44 65 55 120 365 122 19 240 88 115 150 370 16 80 181
[49] 265 300 59 185 360 56 12 70 39 24 32 1000 110 270 22 75 250 62 23 1250 364 74 198 100
[73] 500 43 91 480 53 99 160 47 999 186 366 68 93 87 183 299 175 98 133 354 42 33 37 225
[97] 400 105 184 153 134 222 58 210 275 182 114 85 36
> #Converting neighbourhood_group,neighbourhood and room_type into factors
> airbnb$neighbourhood_group <- as.factor(airbnb$neighbourhood_group)
> levels(airbnb$neighbourhood_group)
[1] "Bronx" "Brooklyn" "Manhattan" "Queens" "Staten Island"
> airbnb$room_type <- as.factor(airbnb$room_type)
> levels(airbnb$room_type)
[1] "Entire home/apt" "Private room" "Shared room"
> airbnb$neighbourhood <- as.factor(airbnb$neighbourhood)
> levels(airbnb$neighbourhood)
[1] "Allerton" "Arden Heights" "Arrochar" "Arverne"
[5] "Astoria" "Bath Beach" "Battery Park City" "Bay Ridge"
[9] "Bay Terrace" "Bay Terrace, Staten Island" "Baychester" "Bayside"
[13] "Bayswater" "Bedford-Stuyvesant" "Belle Harbor" "Bellerose"
[17] "Belmont" "Bensonhurst" "Bergen Beach" "Boerum Hill"
[21] "Borough Park" "Breezy Point" "Briarwood" "Brighton Beach"
[25] "Bronxdale" "Brooklyn Heights" "Brownsville" "Bull's Head"
[29] "Bushwick" "Cambria Heights" "Canarsie" "Carroll Gardens"
Links
ENG 12:36 AM

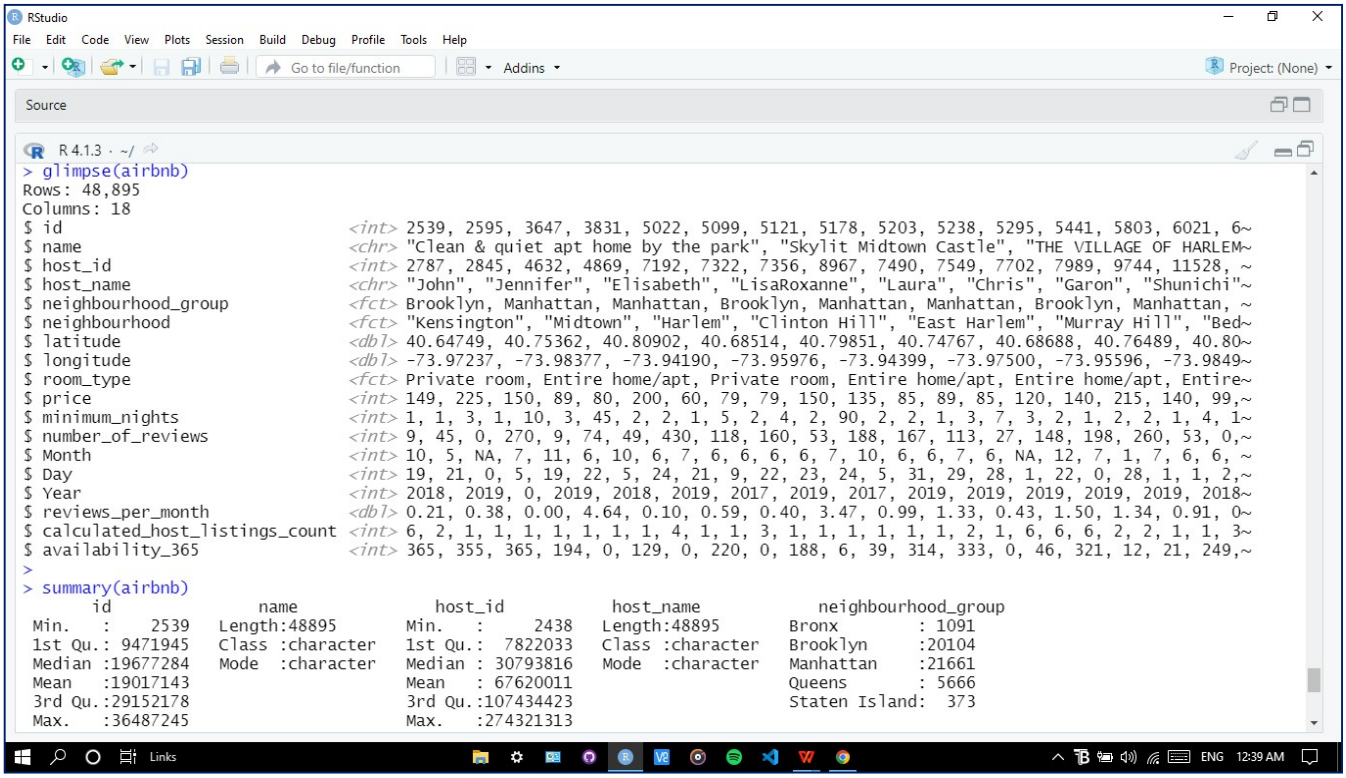

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Source
R 4.1.3 · ~/
> #Converting neighbourhood_group,neighbourhood and room_type into factors
> airbnb$neighbourhood_group <- as.factor(airbnb$neighbourhood_group)
> levels(airbnb$neighbourhood_group)
[1] "Bronx" "Brooklyn" "Manhattan" "Queens" "Staten Island"
> airbnb$room_type <- as.factor(airbnb$room_type)
> levels(airbnb$room_type)
[1] "Entire home/apt" "Private room" "Shared room"
> airbnb$neighbourhood <- as.factor(airbnb$neighbourhood)
> levels(airbnb$neighbourhood)
[1] "Allerton" "Arden Heights" "Arrochar" "Arverne"
[5] "Astoria" "Bath Beach" "Battery Park City" "Bay Ridge"
[9] "Bay Terrace" "Bay Terrace, Staten Island" "Baychester" "Bayside"
[13] "Bayswater" "Bedford-Stuyvesant" "Belle Harbor" "Bellerose"
[17] "Belmont" "Bensonhurst" "Bergen Beach" "Boerum Hill"
[21] "Borough Park" "Breezy Point" "Briarwood" "Brighton Beach"
[25] "Bronxdale" "Brooklyn Heights" "Brownsville" "Bull's Head"
[29] "Bushwick" "Cambria Heights" "Canarsie" "Carroll Gardens"
[33] "Castle Hill" "Castleton Corners" "Chelsea" "Chinatown"
[37] "City Island" "Civic Center" "Claremont Village" "Clason Point"
[41] "Clifton" "Clinton Hill" "Co-op City" "Cobble Hill"
[45] "College Point" "Columbia St" "Concord" "Concourse"
[49] "Concourse Village" "Coney Island" "Corona" "Crown Heights"
[53] "Cypress Hills" "Ditmars Steinway" "Dongan Hills" "Douglaston"
[57] "Downtown Brooklyn" "DUMBO" "Dyker Heights" "East Elmhurst"
[61] "East Flatbush" "East Harlem" "East Morrisania" "East New York"
[65] "East Village" "Eastchester" "Edenwald" "Edgemere"
[69] "Elmhurst" "Eltingville" "Emerson Hill" "Far Rockaway"
Links
ENG 12:37 AM

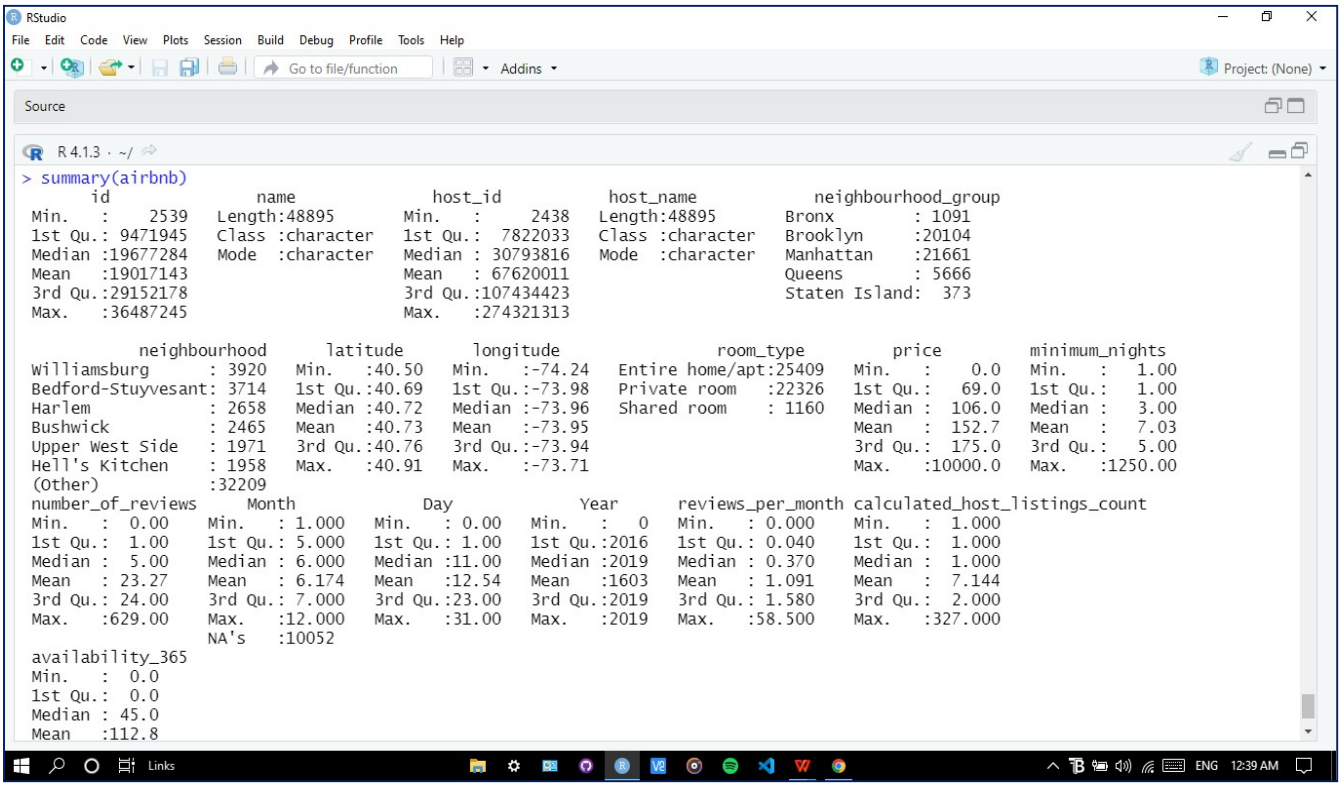


```
> 
> #Separating Dates
> airbnb <- tidyr::separate(airbnb,last_review,c("Month","Day","Year"),sep = "/")
Warning message:
Expected 3 pieces. Missing pieces filled with `NA` in 10052 rows [3, 20, 27, 37, 39, 194, 205, 261, 266, 268, 277, 346, 350, 391,
 426, 433, 438, 487, 546, 586, ...].
> 
> #Replacing NA with 0
> airbnb$Year[is.na(airbnb$Year) == TRUE]=0
> airbnb$Month[is.na(airbnb$Month) == TRUE]=0
> airbnb$Day[is.na(airbnb$Day) == TRUE]=0
> 
> 
> #Datatype Conversion
> airbnb$Month<- as.integer(airbnb$Month)
> airbnb$Year<- as.integer(airbnb$Year)
> airbnb$Day<- as.integer(airbnb$Day)
> View(airbnb)
> 
> 
> #Replacing NA in review_per_month with 0
> airbnb$reviews_per_month[is.na(airbnb$reviews_per_month) == TRUE] = 0
> 
> #Checking for any NA value remaining
> sapply(airbnb, function(x) sum(is.na(x)))
                          id                         name                      host_id                    host_name 
                           0                            0                            0                            0 
           neighbourhood_group                neighbourhood                     latitude                    longitude 
                           0                            0                            0                            0 
                     room_type                        price                minimum_nights            number_of_reviews 
```



```
> #Replacing NA in review_per_month with 0
> airbnb$reviews_per_month[is.na(airbnb$reviews_per_month) == TRUE] = 0
> 
> #Checking for any NA value remaining
> sapply(airbnb, function(x) sum(is.na(x)))
                          id                         name                      host_id                    host_name 
                           0                            0                            0                            0 
           neighbourhood_group                neighbourhood                     latitude                    longitude 
                           0                            0                            0                            0 
                     room_type                        price                minimum_nights            number_of_reviews 
                           0                            0                            0                            0 
                         Month                          Day                         Year              reviews_per_month 
                       10052                            0                            0                            0 
calculated_host_listings_count               availability_365 
                           0                            0 
> glimpse(airbnb)
Rows: 48,895
Columns: 18
$ id                            <int> 2539, 2595, 3647, 3831, 5022, 5099, 5121, 5178, 5203, 5238, 5295, 5441, 5803, 6021, 6~
$ name                          <chr> "Clean & quiet apt home by the park", "Skylit Midtown Castle", "THE VILLAGE OF HARLEM~
$ host_id                       <int> 2787, 2845, 4632, 4869, 7192, 7322, 7356, 8967, 7490, 7549, 7702, 7989, 9744, 11528, ~
$ host_name                     <chr> "John", "Jennifer", "Elisabeth", "LisaRoxanne", "Laura", "Chris", "Garon", "Shunichi"~
$ neighbourhood_group           <fct> Brooklyn, Manhattan, Manhattan, Brooklyn, Manhattan, Manhattan, Brooklyn, Manhattan, ~
$ neighbourhood                 <fct> "Kensington", "Midtown", "Harlem", "Clinton Hill", "East Harlem", "Murray Hill", "Bed~
$ latitude                      <dbl> 40.64749, 40.75362, 40.80902, 40.68514, 40.79851, 40.74767, 40.68688, 40.76489, 40.80~
$ longitude                     <dbl> -73.97237, -73.98377, -73.94190, -73.95976, -73.94399, -73.97500, -73.95596, -73.9849~
$ room_type                     <fct> Private room, Entire home/apt, Private room, Entire home/apt, Entire home/apt, Entire~
$ price                         <int> 149, 225, 150, 89, 80, 200, 60, 79, 79, 150, 135, 85, 89, 85, 120, 140, 215, 140, 99,~
$ minimum_nights                <int> 1, 1, 3, 1, 10, 3, 45, 2, 2, 1, 5, 2, 4, 2, 90, 2, 2, 1, 3, 7, 3, 2, 1, 2, 2, 1, 4, 1~
$ number_of_reviews             <int> 9, 45, 0, 270, 9, 74, 49, 430, 118, 160, 53, 188, 167, 113, 27, 148, 198, 260, 53, 0,~
```

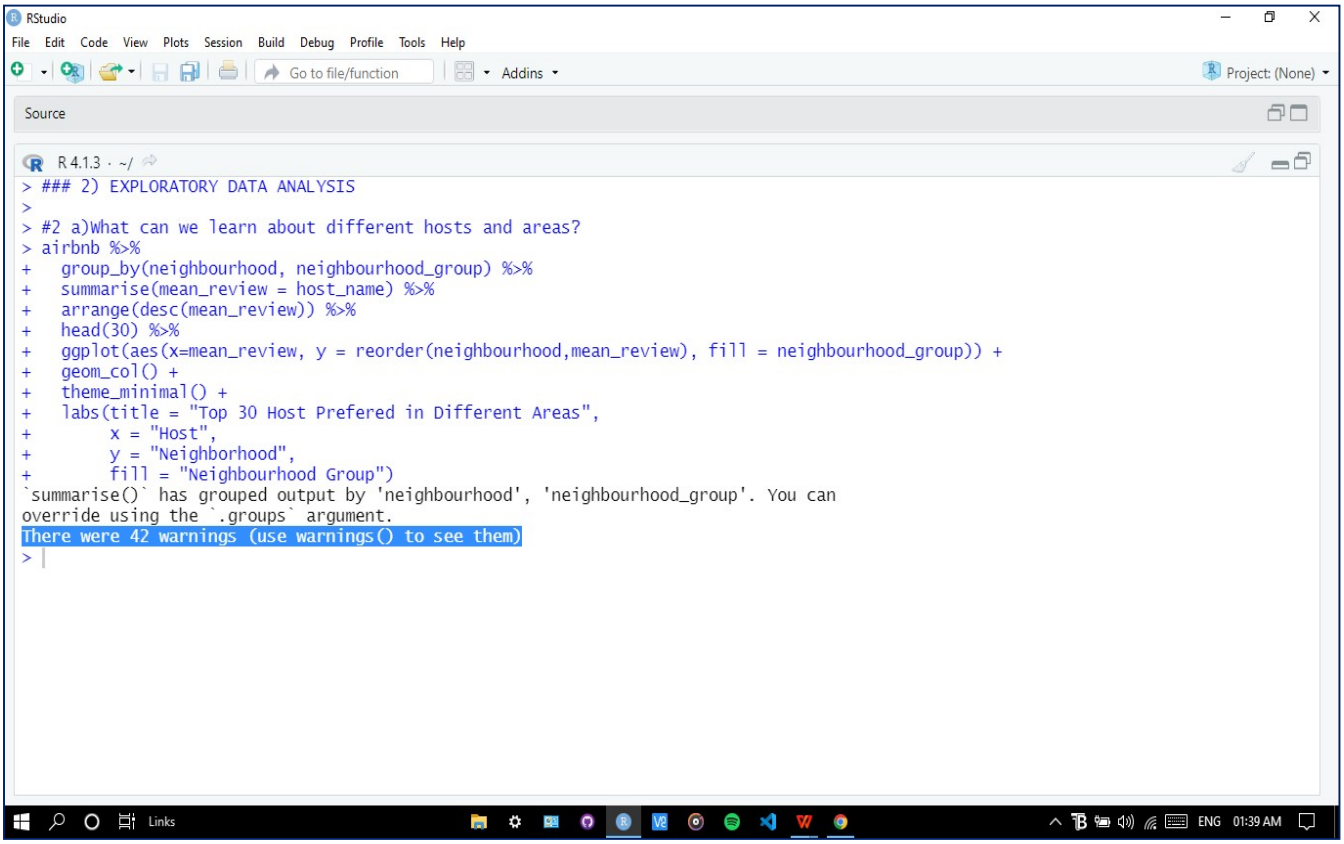

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Source
R 4.1.3 · ~/
> glimpse(airbnb)
Rows: 48,895
Columns: 18
$ id <int> 2539, 2595, 3647, 3831, 5022, 5099, 5121, 5178, 5203, 5238, 5295, 5441, 5803, 6021, 6~
$ name <chr> "Clean & quiet apt home by the park", "Skylit Midtown Castle", "THE VILLAGE OF HARLEM~
$ host_id <int> 2787, 2845, 4632, 4869, 7192, 7322, 7356, 8967, 7490, 7549, 7702, 7989, 9744, 11528, ~
$ host_name <chr> "John", "Jennifer", "Elisabeth", "LisaRoxanne", "Laura", "Chris", "Garon", "Shunichi"~
$ neighbourhood_group <fct> Brooklyn, Manhattan, Manhattan, Brooklyn, Manhattan, Manhattan, Brooklyn, Manhattan, ~
$ neighbourhood <fct> "Kensington", "Midtown", "Harlem", "Clinton Hill", "East Harlem", "Murray Hill", "Bed~
$ latitude <dbl> 40.64749, 40.75362, 40.80902, 40.68514, 40.79851, 40.74767, 40.68688, 40.76489, 40.80~
$ longitude <dbl> -73.97237, -73.98377, -73.94190, -73.95976, -73.94399, -73.97500, -73.95596, -73.9849~
$ room_type <fct> Private room, Entire home/apt, Private room, Entire home/apt, Entire home/apt, Entire~
$ price <int> 149, 225, 150, 89, 80, 200, 60, 79, 79, 150, 135, 85, 89, 85, 120, 140, 215, 140, 99,~
$ minimum_nights <int> 1, 1, 3, 1, 10, 3, 45, 2, 2, 1, 5, 2, 4, 2, 90, 2, 2, 1, 3, 7, 3, 2, 1, 2, 2, 1, 4, 1~
$ number_of_reviews <int> 9, 45, 0, 270, 9, 74, 49, 430, 118, 160, 53, 188, 167, 113, 27, 148, 198, 260, 53, 0,~
$ Month <int> 10, 5, NA, 7, 11, 6, 10, 6, 7, 6, 6, 6, 6, 7, 10, 6, 6, 7, 6, NA, 12, 7, 1, 7, 6, 6, ~
$ Day <int> 19, 21, 0, 5, 19, 22, 5, 24, 21, 9, 22, 23, 24, 5, 31, 29, 28, 1, 22, 0, 28, 1, 1, 2,~
$ Year <int> 2018, 2019, 0, 2019, 2018, 2019, 2017, 2019, 2017, 2019, 2019, 2019, 2019, 2019, 2018~
$ reviews_per_month <dbl> 0.21, 0.38, 0.00, 4.64, 0.10, 0.59, 0.40, 3.47, 0.99, 1.33, 0.43, 1.50, 1.34, 0.91, 0~
$ calculated_host_listings_count <int> 6, 2, 1, 1, 1, 1, 1, 1, 1, 4, 1, 1, 3, 1, 1, 1, 1, 1, 1, 2, 1, 6, 6, 6, 2, 2, 1, 1, 3~
$ availability_365 <int> 365, 355, 365, 194, 0, 129, 0, 220, 0, 188, 6, 39, 314, 333, 0, 46, 321, 12, 21, 249,~
> summary(airbnb)
id name host_id host_name neighbourhood_group
Min. : 2539 Length:48895 Min. : 2438 Length:48895 Bronx : 1091
1st Qu.: 9471945 Class :character 1st Qu.: 7822033 Class :character Brooklyn :20104
Median :19677284 Mode :character Median : 30793816 Mode :character Manhattan :21661
Mean :19017143 Mean : 67620011 Queens : 5666
3rd Qu.:29152178 3rd Qu.:107434423 Staten Island: 373
Max. :36487245 Max. :274321313
Links
ENG 12:39 AM

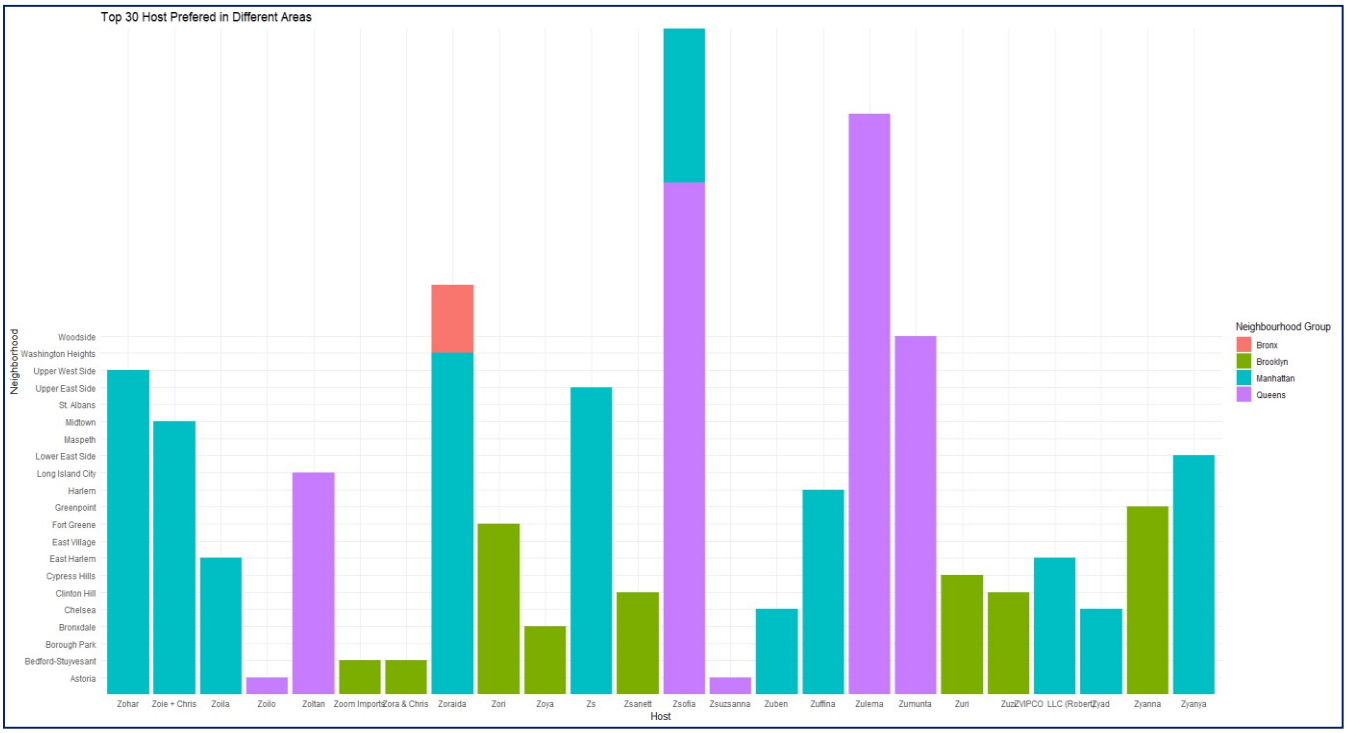

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Source
R 4.1.3 · ~/
> summary(airbnb)
id name host_id host_name neighbourhood_group
Min. : 2539 Length:48895 Min. : 2438 Length:48895 Bronx : 1091
1st Qu.: 9471945 Class :character 1st Qu.: 7822033 Class :character Brooklyn :20104
Median :19677284 Mode :character Median : 30793816 Mode :character Manhattan :21661
Mean :19017143 Mean : 67620011 Queens : 5666
3rd Qu.:29152178 3rd Qu.:107434423 Staten Island: 373
Max. :36487245 Max. :274321313
neighbourhood latitude longitude room_type price minimum_nights
Williamsburg : 3920 Min. :40.50 Min. :-74.24 Entire home/apt:25409 Min. : 0.0 Min. : 1.00
Bedford-Stuyvesant: 3714 1st Qu.:40.69 1st Qu.:-73.98 Private room :22326 1st Qu.: 69.0 1st Qu.: 1.00
Harlem : 2658 Median :40.72 Median :-73.96 Shared room : 1160 Median : 106.0 Median : 3.00
Bushwick : 2465 Mean :40.73 Mean :-73.95 Mean : 152.7 Mean : 7.03
Upper West Side : 1971 3rd Qu.:40.76 3rd Qu.:-73.94 3rd Qu.: 175.0 3rd Qu.: 5.00
Hell's Kitchen : 1958 Max. :40.91 Max. :-73.71 Max. :10000.0 Max. :1250.00
(Other) :32209
number_of_reviews Month Day Year reviews_per_month calculated_host_listings_count
Min. : 0.00 Min. : 1.000 Min. : 0.00 Min. : 0 Min. : 0.000 Min. : 1.000
1st Qu.: 1.00 1st Qu.: 5.000 1st Qu.: 1.00 1st Qu.:2016 1st Qu.: 0.040 1st Qu.: 1.000
Median : 5.00 Median : 6.000 Median :11.00 Median :2019 Median : 0.370 Median : 1.000
Mean : 23.27 Mean : 6.174 Mean :12.54 Mean :1603 Mean : 1.091 Mean : 7.144
3rd Qu.: 24.00 3rd Qu.: 7.000 3rd Qu.:23.00 3rd Qu.:2019 3rd Qu.: 1.580 3rd Qu.: 2.000
Max. :629.00 Max. :12.000 Max. :31.00 Max. :2019 Max. :58.500 Max. :327.000
NA's :10052
availability_365
Min. : 0.0
1st Qu.: 0.0
Median : 45.0
Mean :112.8
Links
ENG 12:39 AM

2) Exploratory Data Analysis :

a) What can we learn about different hosts and areas?

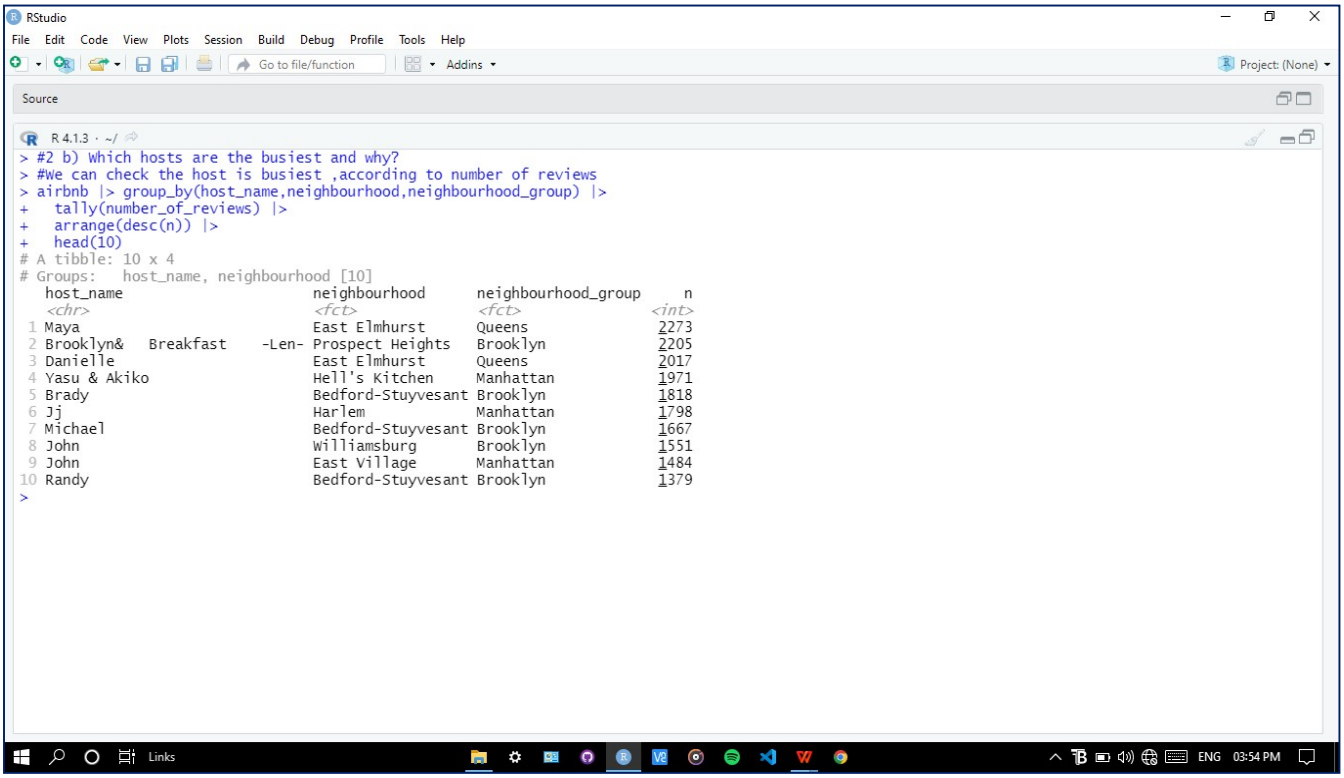
- The graph represents the different areas under different hosts.
- Hosts are stated on the x axis and different areas are shown on the y axis.
- Different colors in the graph represents the different neighborhood groups.
- As per the graph we could say that zsofia host has the more number of area which are queens and manhattan neighborhood groups.
- The Least number of area is under zsuzsanna has the least number of area which is of queens neighborhood groups.





b) Which hosts are the busiest and why?

- **Considering how many reviews it has, we can say the host is busiest as thenumber of reviews increases, therefore we can say the host is busiest as it is trying their best to manage the stuff.**
- **Here Maya, has 2273 reviews, so we can conclude that the host is busiestbecause she is giving her 100% to manage items.**
- **So, Maya is the most busiest host here!**



c) Is there any noticeable difference in traffic among differentareas and what could be the reason for it?

- The graph represents the Top 50 Neighbourhood according to the Number of reviews
- Different colors in the graph represents the different neighborhood groups.
- Yes! There is a lot of difference in traffic among different areas as we can see in the bar plot there is huge traffic in QUEENS EAST ELMHURST as compared to others neighbourhood group.
- From bar plot we can clearly see that the number of reviews of QUEEN EAST ELMHURST is highest ,so this means the number of visits are more in QUEENS EAST ELMHURST so we could conclude that there is more traffic in QUEENS EAST ELMHURST neighbourhood group.
- So, according to number of reviews we can analyse traffic among different areas

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

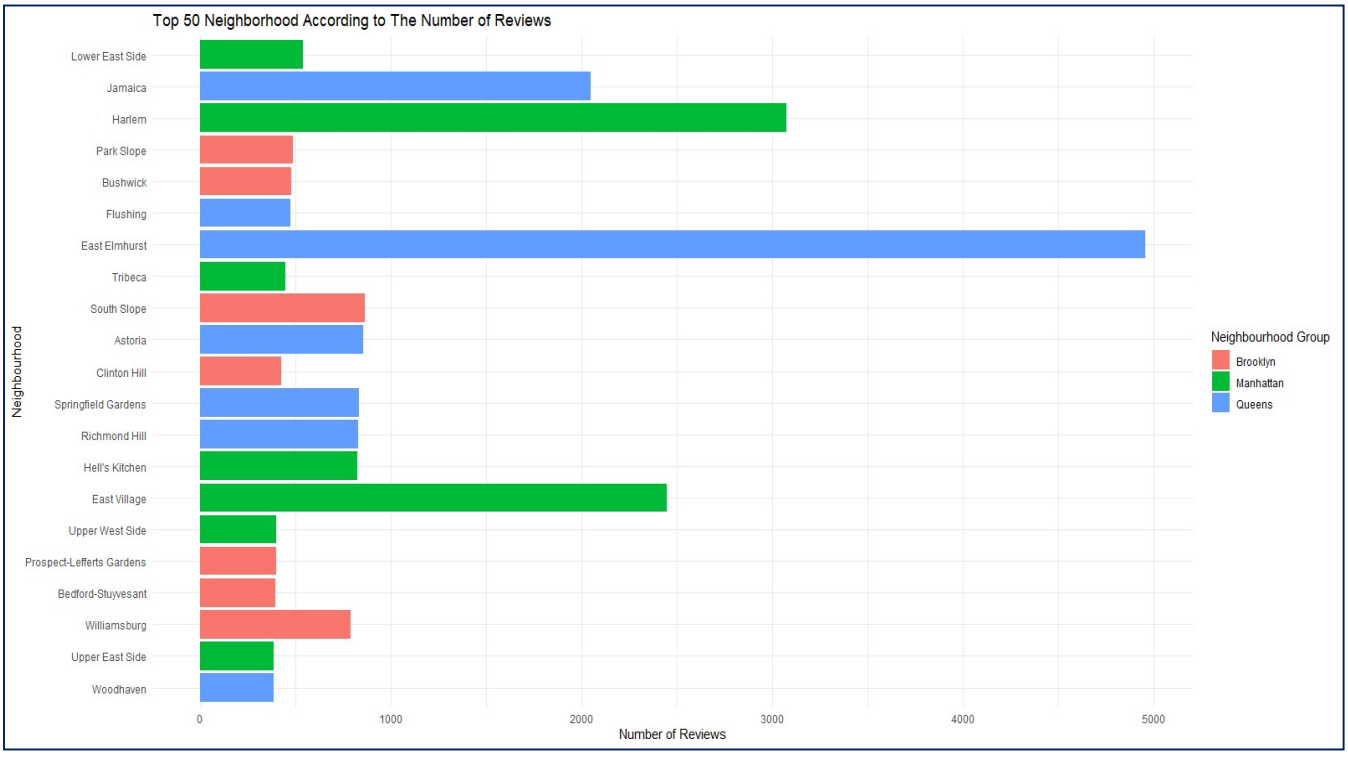
Go to file/function | Addins | Project: (None)

Source

R 4.1.3 · ~/

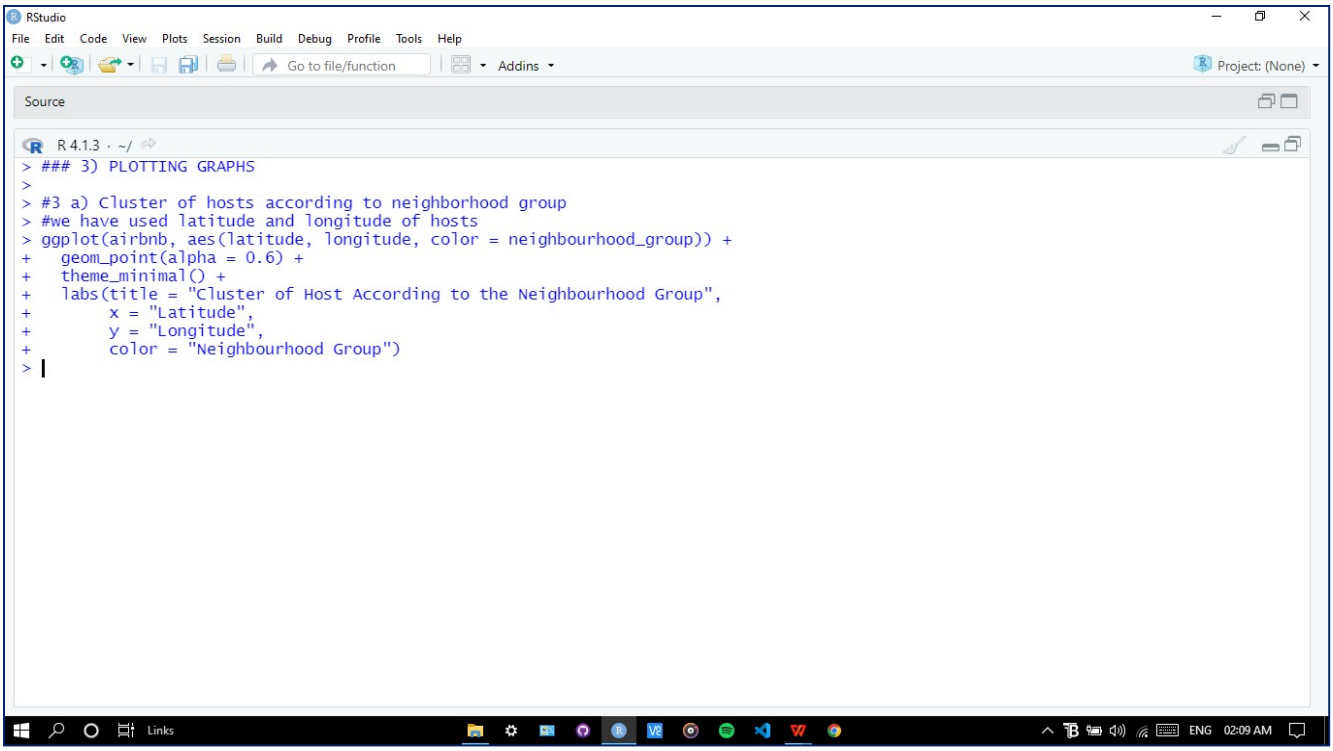
```
> # 2 c)Is there any noticeable difference in traffic among different areas and what could be the reason for it?
> airbnb %>%
+   group_by(neighbourhood, neighbourhood_group) %>%
+   summarise(review = number_of_reviews) %>%
+   arrange(desc(review)) %>%
+   head(50) %>%
+   ggplot(aes(x=review, y = reorder(neighbourhood,review), fill = neighbourhood_group)) +
+   geom_col() +
+   theme_minimal() +
+   labs(title = "Top 50 Neighborhood According to The Number of Reviews",
+        x = "Number of Reviews",
+        y = "Neighbourhood",
+        fill = "Neighbourhood Group")
`summarise()` has grouped output by 'neighbourhood', 'neighbourhood_group'. You can override using the `.groups` argument.
> 
```

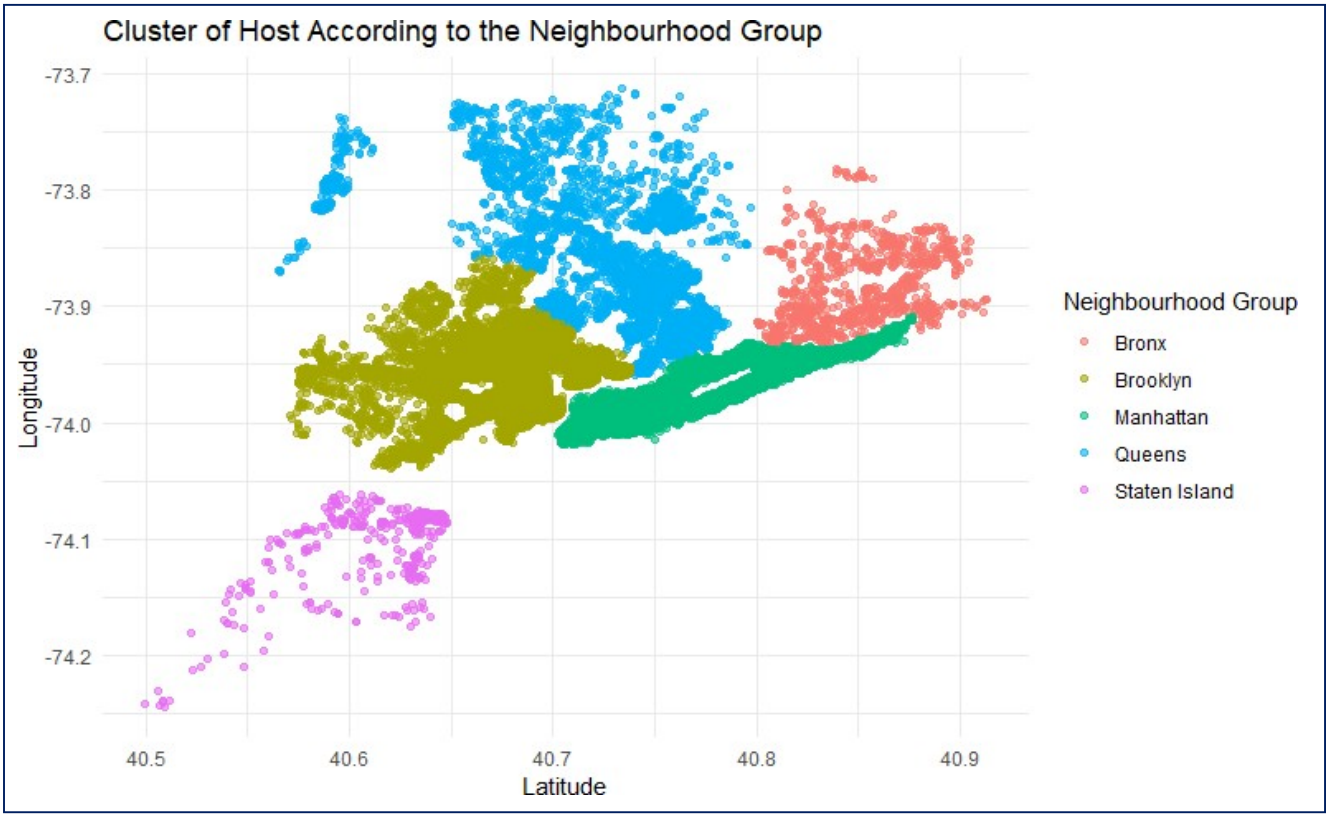
Links 05:55 PM



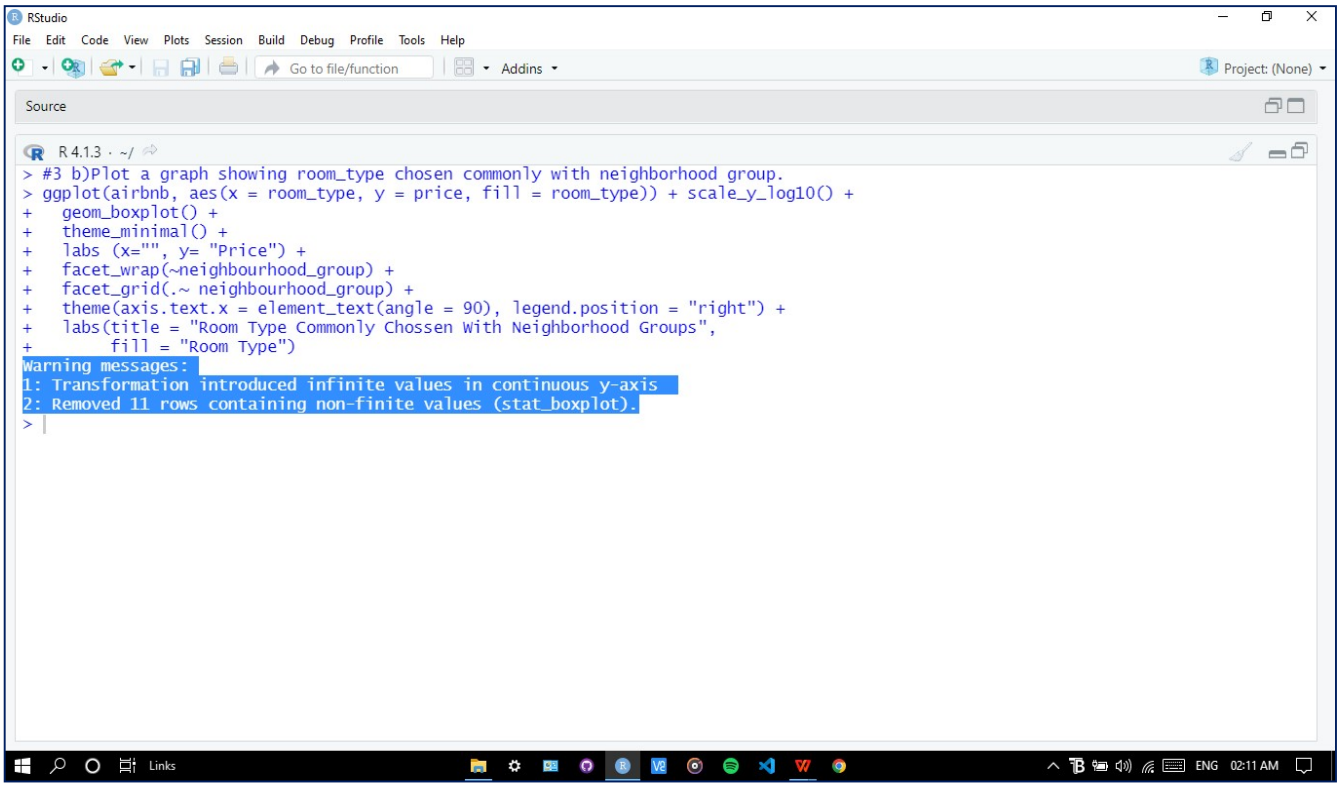
3) Plot Graphs Based On Some Conditions

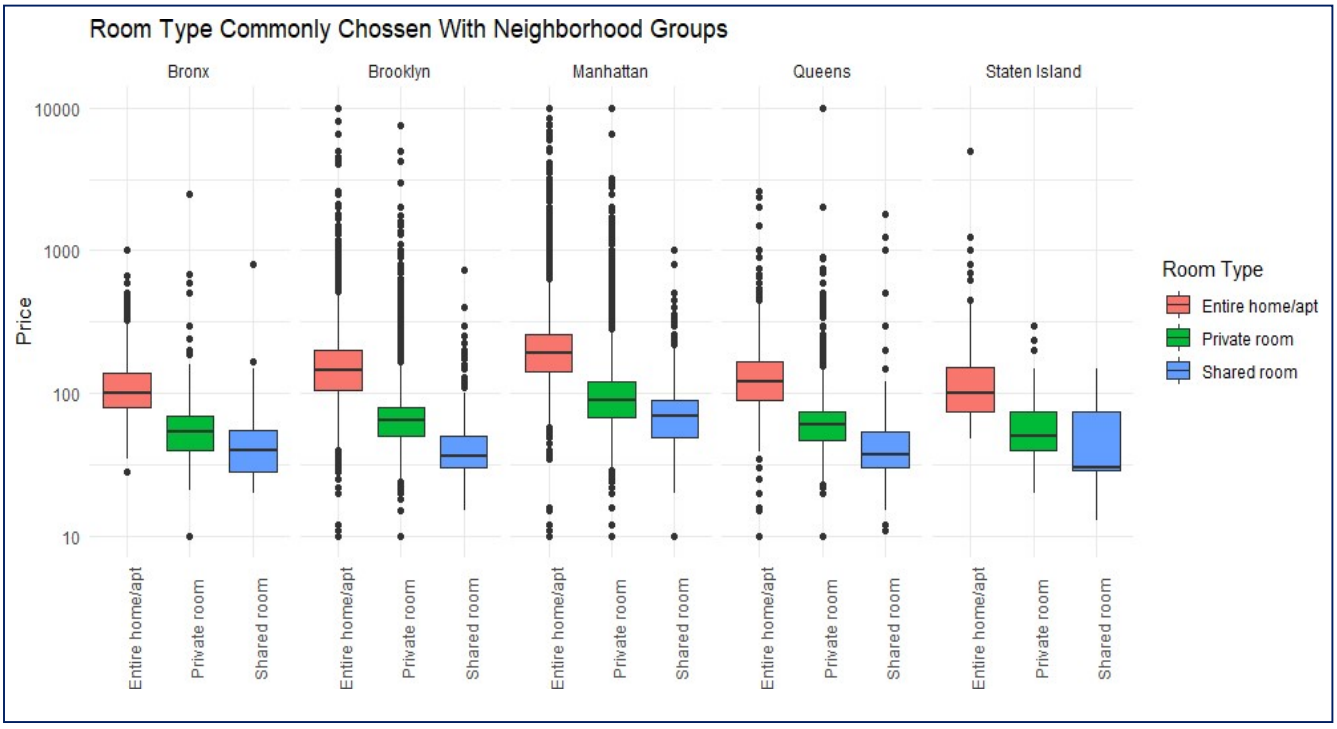
a) Cluster of hosts according to neighborhood group :



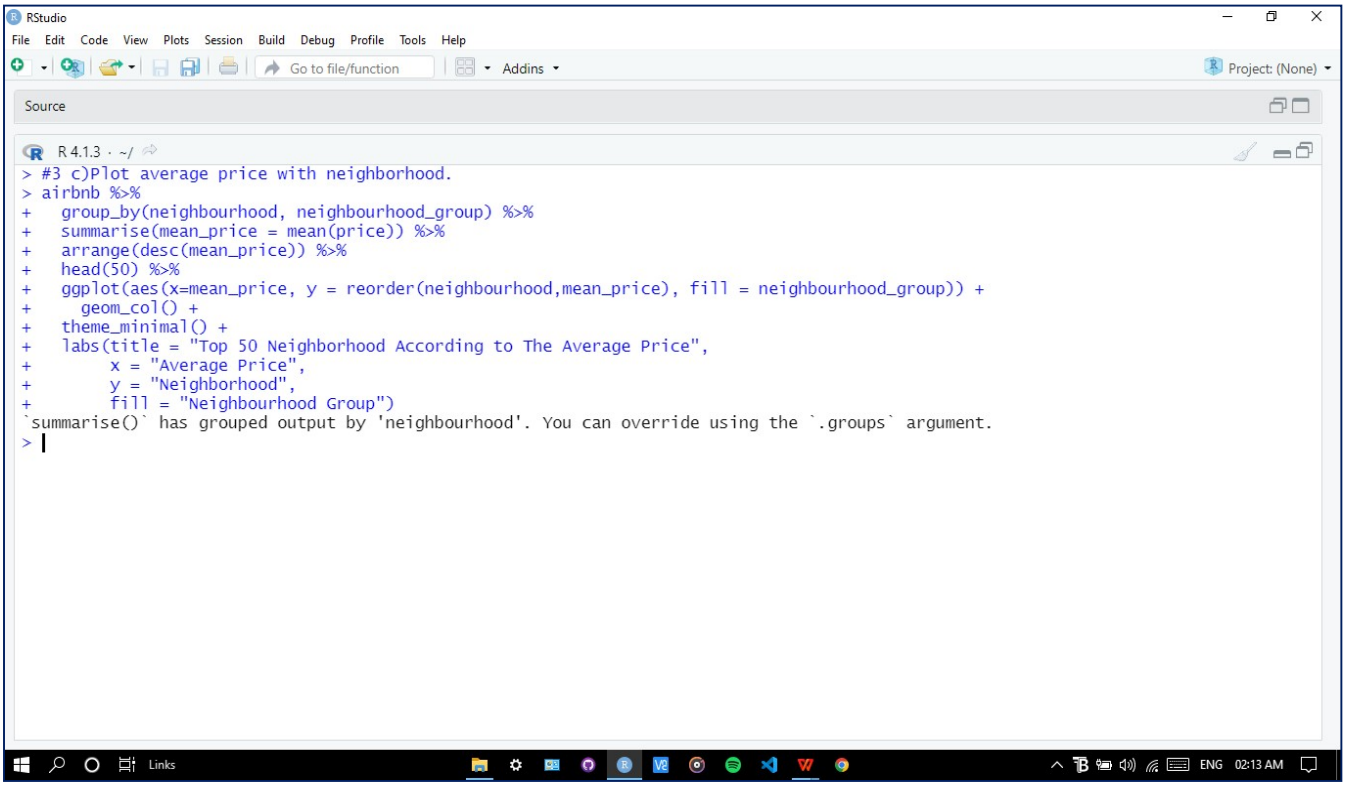


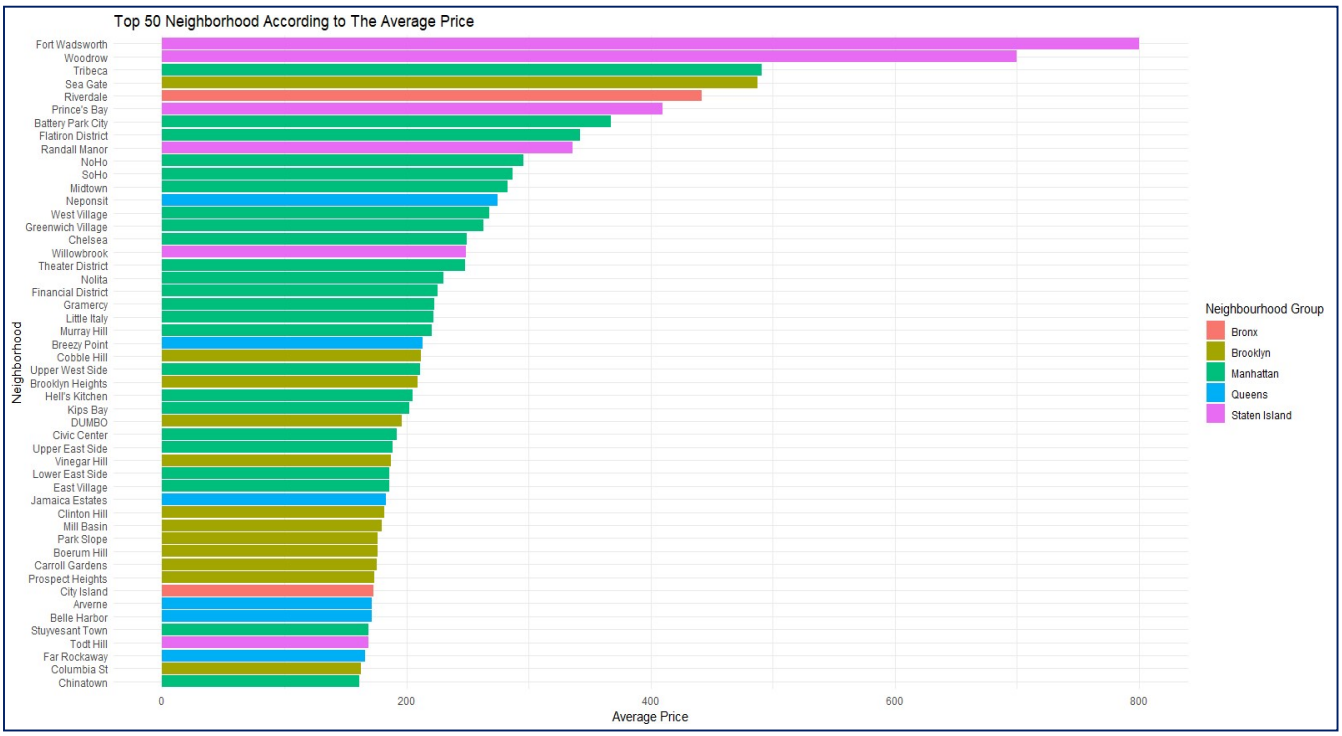
b) Plot a graph showing room_type chosen commonly withneighborhood group :





c) Plot average price with neighborhood :





d) Plot Highest, Lowest, and Median price of stay with neighborhood :

1) HIGHEST PRICE OF STAY WITH NEIGHBORHOOD :

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

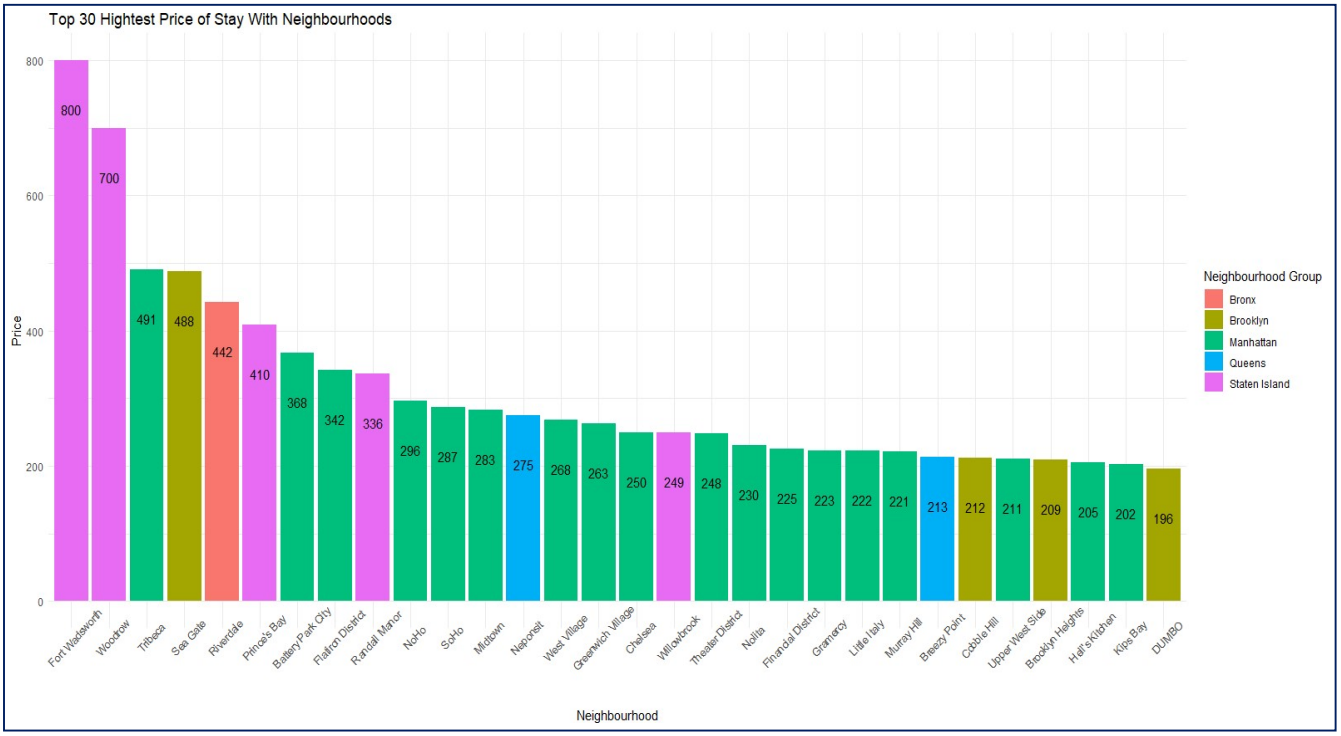
Go to file/function | Addins | Project: (None)

Source

R 4.1.3 · ~/

```
> #3 d)Plot highest, lowest, and median price of stay with neighborhood
> #highest
> airbnb %>%
+   group_by(neighbourhood_group,neighbourhood)%>%
+   summarise(mean_price = mean(price))%>%
+   arrange(desc(mean_price))%>%
+   head(30)%>%
+   ggplot(., aes(x = reorder(neighbourhood, -mean_price) , y = mean_price, fill = neighbourhood_group)) +
+   geom_col() +
+   theme_minimal() +
+   geom_text(aes(label = format(mean_price,digits=3)), size=4, position = position_dodge(0.9),vjust = 5) +
+   theme(axis.text.x = element_text(angle = 45), legend.position = "right") +
+   labs(title = "Top 30 Hightest Price of Stay With Neighbourhoods",
+        x = "Neighbourhood",
+        y = "Price",
+        fill = "Neighbourhood Group")
`summarise()` has grouped output by 'neighbourhood_group'. You can override using the `.groups` argument.
>
```

Links | ENG 12:24 PM



2) LOWEST PRICE OF STAY WITH NEIGHBORHOOD :

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

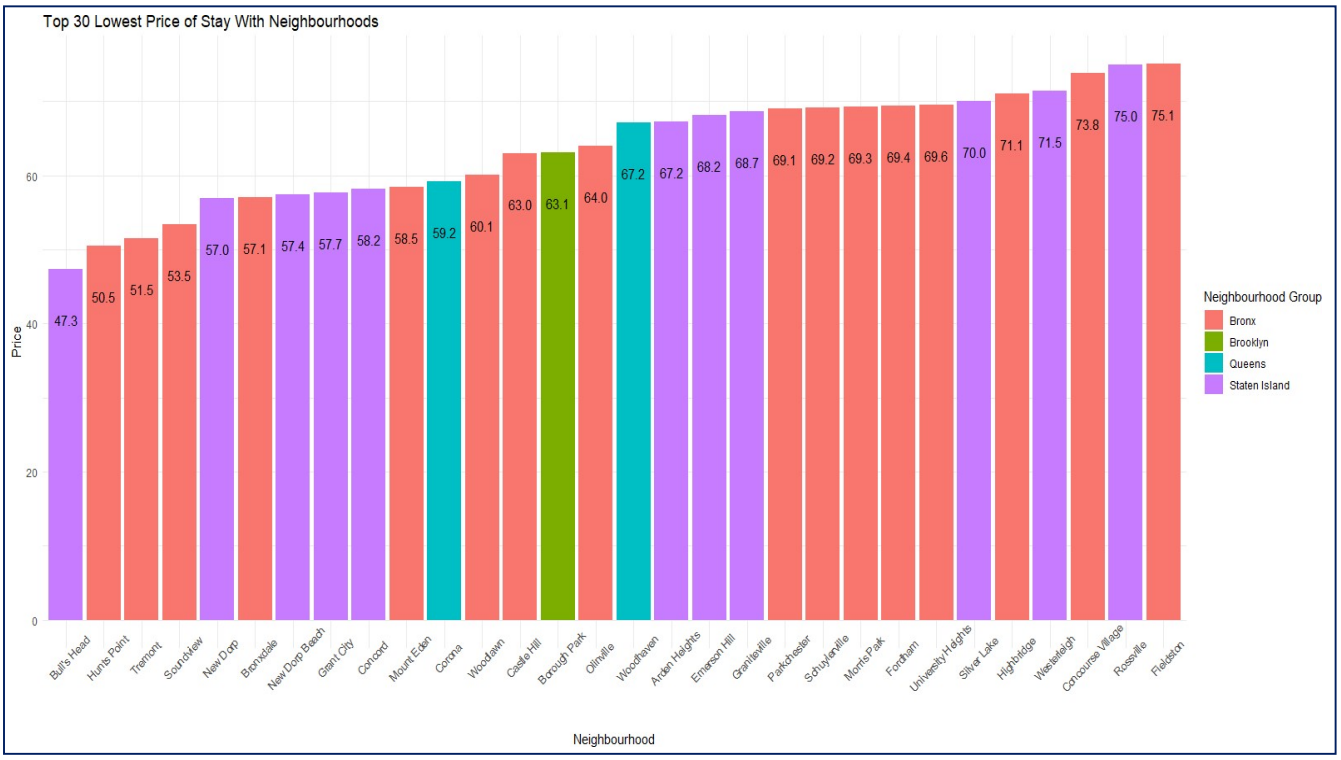
Go to file/function | Addins | Project: (None)

Source

R 4.1.3 · ~/

```
> #lowest
> airbnb %>%
+   group_by(neighbourhood_group,neighbourhood)%>%
+   summarise(mean_price = mean(price))%>%
+   arrange(mean_price) %>%
+   head(30)%>%
+   ggplot(., aes(x = reorder(neighbourhood, mean_price) , y = mean_price, fill = neighbourhood_group)) +
+   geom_col() +
+   theme_minimal() +
+   geom_text(aes(label = format(mean_price,digits=3)), size=4, position = position_dodge(0.9),vjust = 5) +
+   theme(axis.text.x = element_text(angle = 45), legend.position = "right") +
+   labs(title = "Top 30 Lowest Price of Stay With Neighbourhoods",
+        x = "Neighbourhood",
+        y = "Price",
+        fill = "Neighbourhood Group")
`summarise()` has grouped output by 'neighbourhood_group'. You can override
using the `.groups` argument.
>
```

Links | ENG 12:41 PM



3) MEDIAN PRICE OF STAY WITH NEIGHBORHOOD :

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Source

R 4.1.3 · ~/

```
> #median
> airbnb %>%
+   group_by(neighbourhood_group,neighbourhood)%>%
+   summarise(min_price = median(price))%>%
+   arrange(min_price) %>%
+   head(30)%>%
+   ggplot(., aes(x = reorder(neighbourhood, min_price) , y = min_price, fill = neighbourhood_group)) +
+   geom_col() +
+   theme_minimal() +
+   geom_text(aes(label = format(min_price,digits=3)), size=4, position = position_dodge(0.9),vjust = 5) +
+   theme(axis.text.x = element_text(angle = 45), legend.position = "right") +
+   labs(title = "Top 30 Median Price of Stay With Neighbourhoods",
+        x = "Neighbourhood",
+        y = "Price",
+        fill = "Neighbourhood Group")
`summarise()` has grouped output by 'neighbourhood_group'. You can
override using the `.groups` argument.
>
```

Links ENG 12:49 PM