

R PROGRAMMING PROJECT

EXPLORATORY DATA ANALYSIS

REPORT (INFERENCES)

RAASHI BAFNA

PES2UG22CS422

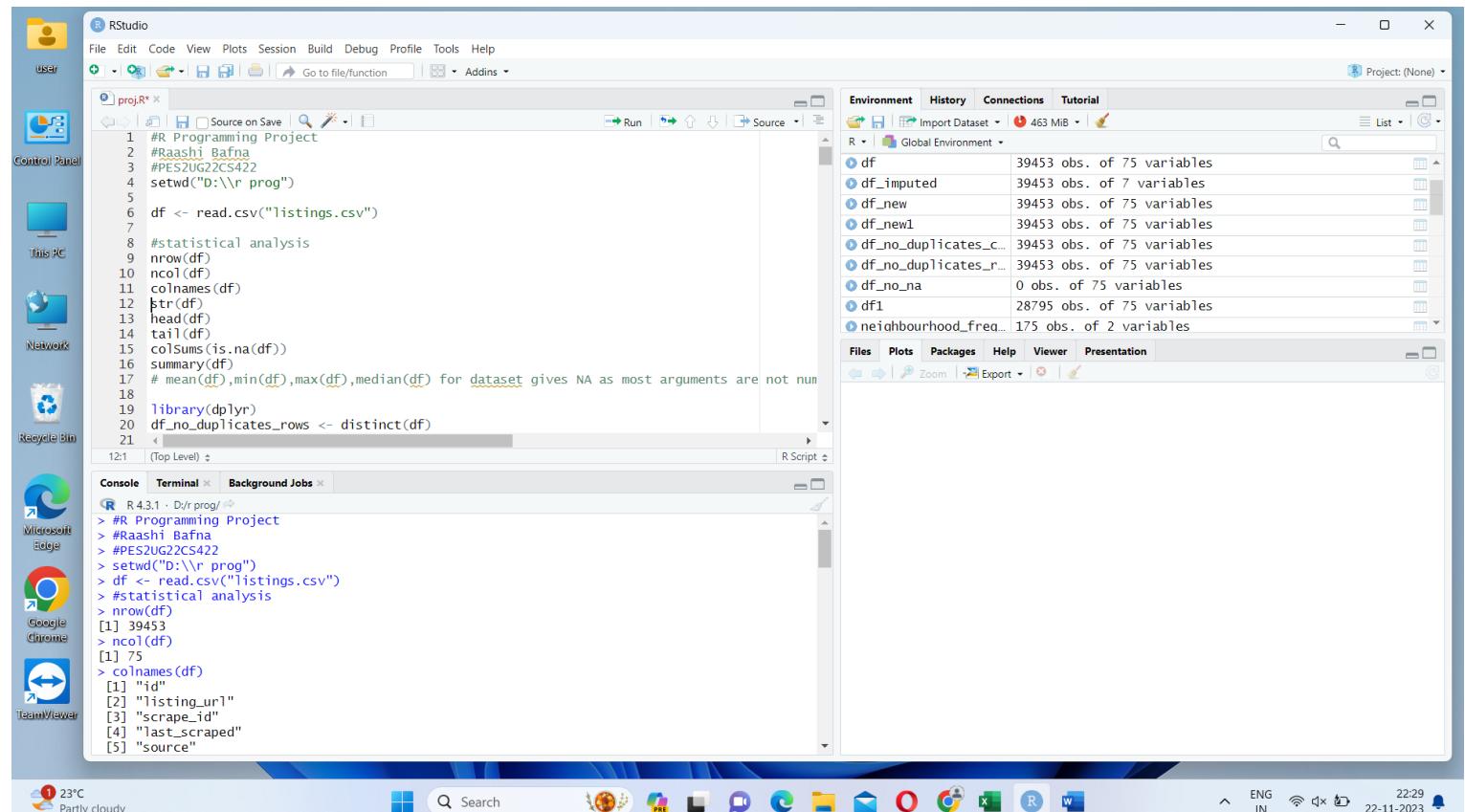
PES2202201371

SEM :3 SEC: G

**FINAL INSIGHTS ARE MENTIONED AT THE END OF THE
FILE(pg.35)**

CODE WITH OUTPUT SCREENSHOTS:

Statistical Analysis



The screenshot shows an RStudio interface with the following details:

- Environment Pane:** Shows the global environment with objects like df, df_imputed, df_new, df_new1, df_no_duplicates_c..., df_no_duplicates_r..., df_no_na, df1, and neighbourhood_freq... along with their respective sizes.
- Code Editor:** A script named "projR.R" containing R code for data analysis, including reading a CSV file, performing statistical analysis, and handling duplicates.
- Console:** Displays the output of the R code, showing the size of the dataset (39453 rows, 75 columns), column names, and the result of the distinct() function.
- System Taskbar:** Shows system icons for Control Panel, Network, Recycle Bin, Microsoft Edge, Google Chrome, and TeamViewer, along with the current weather (23°C Partly cloudy) and system date (22-11-2023).

The dataset has 39453 rows and 75 columns

A screenshot of the RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. A toolbar below has icons for file operations like Open, Save, and Run. The left sidebar shows a project tree with 'proj.Rx' selected, containing an R script with code for reading a CSV file and performing statistical analysis. The right sidebar displays the Environment pane showing various data frames and their dimensions. The bottom pane is the Console, showing the output of running the script, which lists numerous host-related variables. The taskbar at the bottom shows other open applications like Microsoft Edge, Google Chrome, and TeamViewer.

The screenshot shows an RStudio interface running on a Windows operating system. The title bar reads "RStudio". The menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help, and Addins. The top toolbar has icons for file operations like Open, Save, Print, and Go to file/function. The left sidebar shows icons for Control Panel, This PC, Network, and Recycle Bin. The main area consists of several panes:

- Script Editor:** A code editor titled "proj.R x" containing R code for reading a CSV file and performing statistical analysis.
- Environment Browser:** A tree view of the global environment showing objects like df, df_imputed, df_new, etc., with their sizes and variable counts.
- Console:** A list of command-line entries, mostly column names from the dataset, starting with [29] "neighbourhood_cleansed".
- Plots, Packages, Help, Viewer, Presentation:** Buttons for navigating between different RStudio features.

```

R Programming Project
#Raashi_Bafna
#PES2UG22CS422
setwd("D:\\r\\prog")
df <- read.csv("listings.csv")
#statistical analysis
nrow(df)
ncol(df)
colnames(df)
str(df)
head(df)

```

Console Terminal Background Jobs

```

R 4.3.1 - D:/r/prog/ >
[51] "has_availability"
[52] "availability_30"
[53] "availability_60"
[54] "availability_90"
[55] "availability_365"
[56] "calendar_last_scraped"
[57] "number_of_reviews"
[58] "number_of_reviews_ltm"
[59] "number_of_reviews_130d"
[60] "first_review"
[61] "last_review"
[62] "review_scores_rating"
[63] "review_scores_accuracy"
[64] "review_scores_cleanliness"
[65] "review_scores_checkin"
[66] "review_scores_communication"
[67] "review_scores_location"
[68] "review_scores_value"
[69] "license"
[70] "instant_bookable"
[71] "calculated_host_listings_count"
[72] "calculated_host_listings_count_entire_homes"
[73] "calculated_host_listings_count_private_rooms"

```

```

df <- read.csv("listings.csv")
#statistical analysis
nrow(df)
ncol(df)
colnames(df)
str(df)
head(df)
tail(df)
colSums(is.na(df))
summary(df)
# mean(df), min(df), max(df), median(df) for dataset gives NA as most arguments are not numeric or log

```

```

head(df)
  id          listing_url      scrape_id last_scraped      source
1 5.270202e+07 https://www.airbnb.com/rooms/52702018 2.023091e+13 2023-09-05 city scrape
2 9.037761e+17 https://www.airbnb.com/rooms/90377610685399654 2.023091e+13 2023-09-05 city scrape
3 7.849694e+17 https://www.airbnb.com/rooms/784969376930125242 2.023091e+13 2023-09-05 city scrape
4 5.408989e+07 https://www.airbnb.com/rooms/54089885 2.023091e+13 2023-09-05 city scrape
5 8.368432e+17 https://www.airbnb.com/rooms/83684321785604295 2.023091e+13 2023-09-05 city scrape
6 8.411290e+17 https://www.airbnb.com/rooms/841128954402711018 2.023091e+13 2023-09-05 city scrape
  name
1 Rental unit in Queens · ★4.90 · 1 bedroom · 2 beds · 1 bath
2 Rental unit in Queens · ★New · 3 bedrooms · 4 beds · 1 bath
3 Home in Queens · ★4.97 · 3 bedrooms · 4 beds · 1 bath
4 Rental unit in Bronx · ★4.75 · 2 bedrooms · 3 beds · 1 bath
5 Rental unit in The Bronx · ★4.56 · 4 bedrooms · 6 beds · 2 baths
6 Rental unit in The Bronx · ★4.80 · 2 bedrooms · 3 beds · 1 bath
  description
1 Unique entire unit 1-bedroom Apartment in a private house. Conveniently located 3 minutes away from La Gua
rdia Airport and 20 minutes drive to Manhattan. With this warm and charming apartment you will have access

```

When `head(df)` is executed the first few lines of the dataset are displayed all screenshots have not been included as it would be very long

RStudio interface showing the execution of the `tail(df)` command. The code editor contains a script named `proj.R` with the following content:

```
7  
8 #statistical analysis  
9 nrow(df)  
10 ncol(df)  
11 colnames(df)  
12 str(df)  
13 head(df)  
14 tail(df)  
15 colSums(is.na(df))  
16 summary(df)  
17 # mean(df),min(df),max(df),median(df) for dataset gives NA as most arguments are not numeric or log  
18  
19
```

The console output shows the last few rows of the dataset:

```
15:1 (Top Level)   
Console Terminal Background Jobs   
R 4.3.1 - D:/r prog/   
5 0 0 0  
6 reviews_per_month  
1 8.53  
2 NA  
3 4.47  
4 1.43  
5 2.50  
6 2.11  
> tail(df)  
id listing_url scrape_id last_scraped  
39448 2.378194e+07 https://www.airbnb.com/rooms/23781939 2.023091e+13 2023-09-06  
39449 3.390812e+07 https://www.airbnb.com/rooms/33908121 2.023091e+13 2023-09-06  
39450 6.499054e+17 https://www.airbnb.com/rooms/649905426134989324 2.023091e+13 2023-09-06  
39451 8.123634e+17 https://www.airbnb.com/rooms/812363361679077941 2.023091e+13 2023-09-06  
39452 1.244649e+07 https://www.airbnb.com/rooms/12446494 2.023091e+13 2023-09-05  
39453 4.929029e+07 https://www.airbnb.com/rooms/49290285 2.023091e+13 2023-09-06  
source name  
39448 previous scrape Rental unit in New York · 2 bedrooms · 2 beds · 1 bath  
39449 previous scrape Rental unit in Brooklyn · 1 bedroom · 2 beds · 1 shared bath  
39450 previous scrape Rental unit in Brooklyn · ★5.0 · 2 bedrooms · 2 beds · 1 private bath  
39451 city scrape Rental unit in New York · ★5.0 · 1 bedroom · 2 beds · 1.5 baths  
39452 previous scrape Rental unit in New York · 2 bedrooms · 2 beds · 1 bath
```

The environment pane shows the following objects:

- df 39453 obs. of 75 variables
- df_imputed 39453 obs. of 7 variables
- df_new 39453 obs. of 75 variables
- df_new1 39453 obs. of 75 variables
- df_no_duplicates... 39453 obs. of 75 variables
- df_no_duplicates... 39453 obs. of 75 variables
- df_no_na 0 obs. of 75 variables
- df1 28795 obs. of 75 variables
- neighbourhood_fri 175 obs. of 2 variables

when `tail(df)` is executed the last few lines of the dataset are displayed
all screenshots have not been included as it would be very long

RStudio interface showing the execution of the `colSums(is.na(df))` command. The code editor contains a script named `proj.R` with the following content:

```
7  
8 #statistical analysis  
9 nrow(df)  
10 ncol(df)  
11 colnames(df)  
12 str(df)  
13 head(df)  
14 tail(df)  
15 colSums(is.na(df))  
16 summary(df)  
17 # mean(df),min(df),max(df),median(df) for dataset gives NA as most arguments are not numeric or log  
18  
19 library(dplyr)  
20
```

The console output shows the number of missing values per column:

```
16:1 (Top Level)   
Console Terminal Background Jobs   
R 4.3.1 - D:/r prog/   
39449 1 0 0  
39450 1 0 0  
39451 0 0 0  
39452 0 0 0  
39453 1 0 0  
reviews_per_month  
39448 0.03  
39449 NA  
39450 0.20  
39451 2.06  
39452 NA  
39453 0.14  
> colSums(is.na(df))
```

The environment pane shows the following objects:

- df 39453 obs. of 75 variables
- df_imputed 39453 obs. of 7 variables
- df_new 39453 obs. of 75 variables
- df_new1 39453 obs. of 75 variables
- df_no_duplicates... 39453 obs. of 75 variables
- df_no_duplicates... 39453 obs. of 75 variables
- df_no_na 0 obs. of 75 variables
- df1 28795 obs. of 75 variables
- neighbourhood_fri 175 obs. of 2 variables

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

proj.R* | Go to file/function | Addins | Run | Source | Environment History Connections Tutorial | Import Dataset | 93 MiB | Project: (None)

```
13 head(df)
14 tail(df)
15 colSums(is.na(df))
16 summary(df)
17 # mean(df),min(df),max(df),median(df) for dataset gives NA as most arguments are not numeric or log
18
19 library(dplyr)
20
```

16:1 (Top Level) | R Script | Environment | History | Connections | Tutorial | Import Dataset | 93 MiB | Project: (None)

Console Terminal Background Jobs x

R 4.3.1 - D:/r prog/

host_url	host_name
host_since	host_location
host_about	host_response_time
host_response_rate	host_acceptance_rate
host_is_superhost	host_thumbnail_url
host_picture_url	host_neighbourhood
host_listings_count	host_total_listings_count
host_verifications	host_has_profile_pic
host_identity_verified	neighbourhood
neighbourhood_cleansed	neighbourhood_group_cleansed
latitude	longitude
property_type	room_type
accommodates	bathrooms
bathrooms_text	bedrooms

23°C Partly cloudy | Search | ENG IN | 22-11-2023 | 22:45

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

proj.R* | Go to file/function | Addins | Run | Source | Environment History Connections Tutorial | Import Dataset | 95 MiB | Project: (None)

```
13 head(df)
14 tail(df)
15 colSums(is.na(df))
16 summary(df)
17 # mean(df),min(df),max(df),median(df) for dataset gives NA as most arguments are not numeric or log
18
19 library(dplyr)
20
```

16:1 (Top Level) | R Script | Environment | History | Connections | Tutorial | Import Dataset | 95 MiB | Project: (None)

Console Terminal Background Jobs x

R 4.3.1 - D:/r prog/

beds	amenities
price	minimum_nights
maximum_nights	minimum_minimum_nights
maximum_minimum_nights	minimum_maximum_nights
maximum_maximum_nights	minimum_nights_avg_ntm
maximum_nights_avg_ntm	calendar_updated
has_availability	availability_30
availability_60	availability_90
availability_365	calendar_last_scraped
number_of_reviews	number_of_reviews_ltm
number_of_reviews_130d	first_review
last_review	review_scores_rating
review_scores_accuracy	review_scores_cleanliness
review_scores_checkin	review_scores_communication

23°C Partly cloudy | Search | ENG IN | 22-11-2023 | 22:46

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

projR*

```

14 tail(df)
15 colSums(is.na(df))
16 summary(df)
17 # mean(df),min(df),max(df),median(df) for dataset gives NA as most arguments are not numeric or log
18
19 library(dplyr)
20 df_no_duplicates_rows <- distinct(df)
21

```

17:1 (Top Level) R Script

Console Terminal Background Jobs

R 4.3.1 - D:/r prog/

```

review_scores_accuracy          0           review_scores_cleanliness      10241
review_scores_checkin          10654       review_scores_communication    10644
review_scores_location          10658       review_scores_value            10650
license                          10661       instant_bookable              10660
calculated_host_listings_count 39453       calculated_host_listings_count_entire_homes 0
calculated_host_listings_count_private_rooms 0       calculated_host_listings_count_shared_rooms 0
reviews_per_month                10241

```

```

> summary(df)
   id          listing_url      scrape_id      last_scraped      source
Min. :2.595e+03 Length:39453 Min. :2.023e+13 Length:39453 Min. : 2234
1st Qu.:1.986e+07 Class :character 1st Qu.:2.023e+13 Class :character 1st Qu.: 16542425
Median :4.543e+07 Mode  :character Median :2.023e+13 Mode  :character Median : 77725998
Mean   :2.778e+17          Mean   :2.023e+13          Mean   :2.023e+13          Mean   :2.023e+13
3rd Qu.:7.135e+17          3rd Qu.:2.023e+13          3rd Qu.:2.023e+13          3rd Qu.:2.023e+13
Max.  :9.733e+17          Max.  :2.023e+13          Max.  :2.023e+13          Max.  :2.023e+13

   name          description      neighborhood_overview      picture_url      host_id
Length:39453 Length:39453 Length:39453 Length:39453 Min. : 2234
Class :character Class :character Class :character Class :character 1st Qu.: 16542425
Mode  :character Mode  :character Mode  :character Mode  :character Median : 77725998

```

23°C Partly cloudy

22:47 22-11-2023

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

projR*

```

14 tail(df)
15 colSums(is.na(df))
16 summary(df)
17 # mean(df),min(df),max(df),median(df) for dataset gives NA as most arguments are not numeric or log
18
19 library(dplyr)
20 df_no_duplicates_rows <- distinct(df)
21

```

17:1 (Top Level) R Script

Console Terminal Background Jobs

R 4.3.1 - D:/r prog/

```

host_url          host_name      host_since      host_location      host_about
Length:39453 Length:39453 Length:39453 Length:39453 Length:39453
Class :character Class :character Class :character Class :character Class :character
Mode  :character Mode  :character Mode  :character Mode  :character Mode  :character

host_response_time host_response_rate host_acceptance_rate host_is_superhost host_thumbnail_url
Length:39453 Length:39453 Length:39453 Length:39453 Length:39453
Class :character Class :character Class :character Class :character Class :character
Mode  :character Mode  :character Mode  :character Mode  :character Mode  :character

host_picture_url host_neighbourhood host_listings_count host_total_listings_count
Length:39453 Length:39453 Min. : 0.0 Min. : 1.0
Class :character Class :character 1st Qu.: 1.0 1st Qu.: 1.0
Mode  :character Mode  :character Median : 2.0 Median : 3.0
                           Mean : 138.5 Mean : 225.3
                           3rd Qu.: 7.0 3rd Qu.: 10.0
                           Max. :4577.0 Max. :9176.0
                           NA's :5 NA's :5

host_verifications host_has_profile_pic host_identity_verified neighbourhood
Length:39453 Length:39453 Length:39453 Length:39453
Class :character Class :character Class :character Class :character
Mode  :character Mode  :character Mode  :character Mode  :character

```

23°C Partly cloudy

22:47 22-11-2023

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

proj.R*

```
14 tail(df)
15 colSums(is.na(df))
16 summary(df)
17 # mean(df),min(df),max(df),median(df) for dataset gives NA as most arguments are not numeric or log
18
19 library(dplyr)
20 df_no_duplicates_rows <- distinct(df)
21
```

17:1 (Top Level) R Script

Console Terminal Background Jobs

R 4.3.1 · D:/r/prog/

neighbourhood_cleansed	neighbourhood_group_cleansed	latitude	longitude
Length:39453	Length:39453	Min. :40.50	Min. :-74.25
Class :character	Class :character	1st Qu.:40.69	1st Qu.:-73.98
Mode :character	Mode :character	Median :40.73	Median :-73.95
		Mean :40.73	Mean :-73.95
		3rd Qu.:40.76	3rd Qu.:-73.93
		Max. :40.91	Max. :-73.71

property_type	room_type	accommodates	bathrooms	bathrooms_text
Length:39453	Length:39453	Min. : 1.000	Mode:logical	Length:39453
Class :character	Class :character	1st Qu.: 2.000	NA's:39453	Class :character
Mode :character	Mode :character	Median : 2.000		Mode :character
		Mean : 2.934		
		3rd Qu.: 4.000		
		Max. :16.000		

bedrooms	beds	amenities	price	minimum_nights
Min. : 1.000	Min. : 1.000	Length:39453	Length:39453	Min. : 1.00
1st Qu.: 1.000	1st Qu.: 1.000	Class :character	Class :character	1st Qu.: 30.00
Median : 1.000	Median : 1.000	Mode :character	Mode :character	Median : 30.00
Mean : 1.597	Mean : 1.658			Mean : 28.05
3rd Qu.: 2.000	3rd Qu.: 2.000			3rd Qu.: 30.00
Max. :50.000	Max. :42.000			Max. :1250.00
NA's :16893	NA's :605			

maximum_nights	minimum_nights	maximum_nights	minimum_nights	maximum_nights
Min. :1.000e+00	Min. : 1.00	Min. : 1.00	Min. :1.000e+00	Min. : 1.000e+00
1st Qu.:1.200e+02	1st Qu.: 30.00	1st Qu.: 30.00	1st Qu.:3.650e+02	1st Qu.: 3.650e+02
Median :3.650e+02	Median : 30.00	Median : 30.00	Median :1.125e+03	Median : 1.125e+03
Mean :1.500e+02	Mean : 32.62	Mean : 32.62	Mean :5.640e+02	Mean : 5.640e+02

23°C Partly cloudy 22:48 22-11-2023

Environment History Connections Tutorial

R Global Environment

df 39453 obs. of 75 variables
df_imputed 39453 obs. of 7 variables
df_new 39453 obs. of 75 variables
df_new1 39453 obs. of 75 variables
df_no_duplicates... 39453 obs. of 75 variables
df_no_duplicates.. 39453 obs. of 75 variables
df_no_na 0 obs. of 75 variables
df1 28795 obs. of 75 variables
neighbourhood_fr 175 obs. of 2 variables

Files Plots Packages Help Viewer Presentation

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

proj.R*

```
14 tail(df)
15 colSums(is.na(df))
16 summary(df)
17 # mean(df),min(df),max(df),median(df) for dataset gives NA as most arguments are not numeric or log
18
19 library(dplyr)
20 df_no_duplicates_rows <- distinct(df)
21
```

17:1 (Top Level) R Script

Console Terminal Background Jobs

R 4.3.1 · D:/r/prog/

meantdi	mednodi	meanuvi	mediuvi	meanuvir	mediuvir
Mean :5.606e+04	Mean : 29.63	Mean : 34.95	Mean :6.549e+05	Mean :1.123e+03	Mean :1.123e+03
3rd Qu.:1.125e+03	3rd Qu.: 30.00	3rd Qu.: 30.00	3rd Qu.:1.125e+03	3rd Qu.:1.125e+03	3rd Qu.:1.125e+03
Max. :2.147e+09	Max. :1250.00	Max. :1250.00	Max. :2.147e+09	Max. :2.147e+09	Max. :2.147e+09

maximum_maximum_nights	minimum_nights_avg_ntm	maximum_nights_avg_ntm	calendar_updated
Min. :1.000e+00	Min. : 1.00	Min. :1.000e+00	Mode:logical
1st Qu.:3.650e+02	1st Qu.: 30.00	1st Qu.:3.650e+02	NA's:39453
Median :1.125e+03	Median : 30.00	Median :1.125e+03	
Mean :1.199e+06	Mean : 34.33	Mean :1.036e+06	
3rd Qu.:1.125e+03	3rd Qu.: 30.00	3rd Qu.:1.125e+03	
Max. :2.147e+09	Max. :1250.00	Max. :2.147e+09	

has_availability	availability_30	availability_60	availability_90	availability_365
Length:39453	Min. : 0.000	Min. : 0.00	Min. : 0.0	Min. : 0.0
Class :character	1st Qu.: 0.000	1st Qu.: 0.00	1st Qu.: 0.0	1st Qu.: 0.0
Mode :character	Median : 0.000	Median : 4.00	Median :16.0	Median :107.0
	Mean : 6.923	Mean :17.22	Mean :30.3	Mean :144.8
	3rd Qu.:11.000	3rd Qu.:34.00	3rd Qu.:61.0	3rd Qu.:291.0
	Max. :30.000	Max. :60.00	Max. :90.0	Max. :365.0

calendar_last_scraped	number_of_reviews	number_of_reviews_ltm	number_of_reviews_130d
Length:39453	Min. : 0.00	Min. : 0.00	Min. : 0.0000
Class :character	1st Qu.: 0.00	1st Qu.: 0.00	1st Qu.: 0.0000
Mode :character	Median : 5.00	Median : 0.00	Median : 0.0000
	Mean : 25.84	Mean : 7.36	Mean : 0.6207
	3rd Qu.: 24.00	3rd Qu.: 6.00	3rd Qu.: 0.0000
	Max. :1834.00	Max. :686.00	Max. :97.0000

23°C Partly cloudy 22:48 22-11-2023

Environment History Connections Tutorial

R Global Environment

df 39453 obs. of 75 variables
df_imputed 39453 obs. of 7 variables
df_new 39453 obs. of 75 variables
df_new1 39453 obs. of 75 variables
df_no_duplicates... 39453 obs. of 75 variables
df_no_duplicates.. 39453 obs. of 75 variables
df_no_na 0 obs. of 75 variables
df1 28795 obs. of 75 variables
neighbourhood_fr 175 obs. of 2 variables

Files Plots Packages Help Viewer Presentation

RStudio

```

14 tail(df)
15 colSums(is.na(df))
16 summary(df)
17 # mean(df),min(df),max(df),median(df) for dataset gives NA as most arguments are not numeric or log
18
19 library(dplyr)
20 df_no_duplicates_rows <- distinct(df)
21
22 <-
17:1 (Top Level) R Script : Environment History Connections Tutorial
R 4.3.1 - D:/r/prog/ proj.R* R Scripts Global Environment
  df 39453 obs. of 75 variables
  df_imputed 39453 obs. of 7 variables
  df_new 39453 obs. of 75 variables
  df_new1 39453 obs. of 75 variables
  df_no_duplicates_... 39453 obs. of 75 variables
  df_no_duplicates... 39453 obs. of 75 variables
  df_no_na 0 obs. of 75 variables
  df1 28795 obs. of 75 variables
  neighbourhood.fr 175 obs. of 2 variables
  
```

Console Terminal Background Jobs

R 4.3.1 - D:/r/prog/ proj.R

```

review_scores_cleanliness review_scores_checkin review_scores_communication review_scores_location
Min. :0.000 Min. :0.000 Min. :0.000 Min. :0.000
1st Qu.:4.500 1st Qu.:4.800 1st Qu.:4.800 1st Qu.:4.620
Median :4.800 Median :4.940 Median :4.960 Median :4.830
Mean :4.627 Mean :4.813 Mean :4.809 Mean :4.725
3rd Qu.:5.000 3rd Qu.:5.000 3rd Qu.:5.000 3rd Qu.:5.000
Max. :5.000 Max. :5.000 Max. :5.000 Max. :5.000
NA's :10644 NA's :10658 NA's :10650 NA's :10661
review_scores_value license instant_bookable calculated_host_listings_count
Min. :0.000 Mode:logical Length:39453 Min. : 1.00
1st Qu.:4.520 NA's:39453 Class:character 1st Qu.: 1.00
Median :4.750 Mode :character Median : 1.00
Mean :4.627 Mean : 1.00
3rd Qu.:4.930 3rd Qu.: 5.00
Max. :5.000 Max. :597.00
NA's :10660
calculated_host_listings_count_entire_homes calculated_host_listings_count_private_rooms
Min. : 0.00 Min. : 0.00
1st Qu.: 0.00 1st Qu.: 0.00
Median : 1.00 Median : 0.00
Mean : 17.37 Mean : 19.98
3rd Qu.: 2.00 3rd Qu.: 2.00
Max. :597.00 Max. :519.00
calculated_host_listings_count_shared_rooms reviews_per_month
Min. : 0.00000 Min. : 0.010
1st Qu.: 0.00000 1st Qu.: 0.120
Median : 0.00000 Median : 0.450

```

23°C Partly cloudy 22-11-2023

RStudio

```

File Edit Code View Plots Session Build Debug Profile Tools Help
User Control Panel This PC Network Recycle Bin Microsoft Edge Google Chrome TeamViewer
Go to file/function Addins
proj.R* Source on Save Run Up Down Source Environment History Connections Tutorial
R 4.3.1 - D:/r/prog/ proj.R
  Summary() R Scripts Global Environment
  mean_of_review_s... 4.81340059038027
  median_of_bedroo... 1
  median_of_host_1... 2
  median_of_host_t... 3
  median_of_review... 4.94
  numeric_vars Named logi [1:75] TRUE FALSE TRUE FALSE FALSE...
  s 'summaryDefault' Named num [1:6] 1 30 30 28.1...
  sd_of_bedrooms 0.976922305138487
  sd_of_host_listi... 620.086962405847
  sd_of_host_total 978.242964791906
  
```

Console Terminal Background Jobs

R 4.3.1 - D:/r/prog/ proj.R

```

calculated_host_listings_count_shared_rooms reviews_per_month
Min. : 0.00000 Min. : 0.010
1st Qu.: 0.00000 1st Qu.: 0.120
Median : 0.00000 Median : 0.450
Mean : 0.04111 Mean : 1.138
3rd Qu.: 0.00000 3rd Qu.: 1.620
Max. :11.00000 Max. :79.820
NA's :10241
> library(dplyr)
> df_no_duplicates_rows <- distinct(df)
> df_no_duplicates_columns <- df[, !duplicated(names(df))]
> nrow(df)
[1] 39453
> ncol(df)
[1] 75
> # to print mean, median and std dev of some columns with int data type after removing duplicates
> #1
> mean_of_host_listing_count <- mean(df$host_listings_count, na.rm = TRUE)
> median_of_host_listing_count <- median(df$host_listings_count, na.rm = TRUE)
> sd_of_host_listing_count <- sd(df$host_listings_count, na.rm = TRUE)

```

23°C Partly cloudy 22-11-2023

The screenshot shows the RStudio interface with the following details:

- File Menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Toolbar:** Includes icons for file operations like Open, Save, Print, and a search bar labeled "Go to file/function".
- Environment Tab:** Shows the Global Environment with variables like `mean_of_review_s...`, `median_of_bedroom...`, etc., and their values.
- Console Tab:** Displays the R session history with commands and their outputs.
- Code Editor:** Shows R script code for calculating mean, median, and standard deviation for host listing counts.
- Plots Tab:** Not visible in the screenshot.
- Packages Tab:** Not visible in the screenshot.
- Help Tab:** Not visible in the screenshot.
- Viewer Tab:** Not visible in the screenshot.
- Presentation Tab:** Not visible in the screenshot.

Code in Editor:

```
# to print mean, median and std dev of some columns with int data type after removing duplicates
#1
mean_of_host_listing_count <- mean(df$host_listings_count, na.rm = TRUE)
median_of_host_listing_count <- median(df$host_listings_count, na.rm = TRUE)
sd_of_host_listing_count <- sd(df$host_listings_count, na.rm = TRUE)

cat("Mean of host_listings_count:", mean_of_host_listing_count, "\n")
cat("Median of host_listings_count:", median_of_host_listing_count, "\n")
cat("Standard Deviation of host_listings_count:", sd_of_host_listing_count, "\n")

#2
mean_of_host_total_listing_count <- mean(df$host_total_listings_count, na.rm = TRUE)
median_of_host_total_listing_count <- median(df$host_total_listings_count, na.rm = TRUE)
sd_of_host_total_listing_count <- sd(df$host_total_listings_count, na.rm = TRUE)
```

Console Output:

```
R 4.3.1 - D:\r\prog\r
> # to print mean, median and std dev of some columns with int data type after removing duplicates
> #1
> mean_of_host_listing_count <- mean(df$host_listings_count, na.rm = TRUE)
> median_of_host_listing_count <- median(df$host_listings_count, na.rm = TRUE)
> sd_of_host_listing_count <- sd(df$host_listings_count, na.rm = TRUE)
> cat("Mean of host_listings_count:", mean_of_host_listing_count, "\n")
Mean of host_listings_count: 138.5023
> cat("Median of host_listings_count:", median_of_host_listing_count, "\n")
Median of host_listings_count: 2
> cat("Standard Deviation of host_listings_count:", sd_of_host_listing_count, "\n")
Standard Deviation of host_listings_count: 620.087
> #2
> mean_of_host_total_listing_count <- mean(df$host_total_listings_count, na.rm = TRUE)
> median_of_host_total_listing_count <- median(df$host_total_listings_count, na.rm = TRUE)
> sd_of_host_total_listing_count <- sd(df$host_total_listings_count, na.rm = TRUE)
> cat("Mean of host_total_listing_count:", mean_of_host_total_listing_count, "\n")
Mean of host_total_listing_count: 225.324
> cat("Median of host_total_listing_count:", median_of_host_total_listing_count, "\n")
Median of host_total_listing_count: 3
> cat("Standard Deviation of host_total_listing_count:", sd_of_host_total_listing_count, "\n")
```

The screenshot shows the RStudio interface with the following details:

- File Menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Toolbar:** Source on Save, Run, Source, Addins.
- Code Editor:** A script named "proj.R" containing R code for calculating statistics like mean, median, and standard deviation for host total listing count and bedrooms.
- Environment Pane:** Shows the global environment with variables like `mean_of_host_listings`, `median_of_host_total_listing_count`, etc., and their values.
- Console:** Displays the R session output, showing the results of the calculations.
- Taskbar:** Includes icons for File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help, and a search bar.

RStudio

File Edt Code View Plots Session Build Debug Profile Tools Help

proj.R*

```

55 #A
56 mean_of_review_scores_checkin <- mean(df$review_scores_checkin , na.rm = TRUE)
57 median_of_review_scores_checkin <- median(df$review_scores_checkin , na.rm = TRUE)
58 sd_of_review_scores_checkin <- sd(df$review_scores_checkin , na.rm = TRUE)
59
60 cat("Mean of bedrooms:", mean_of_review_scores_checkin , "\n")
61 cat("Median of bedrooms:", median_of_review_scores_checkin , "\n")
62 cat("Standard Deviation of bedrooms:", sd_of_review_scores_checkin , "\n")
63
64 #if all rows with atleast one na value is removed
65 df_no_na <- na.omit(df)
66 nrow(df_no_na)
67 #will be zero as column license column has all na values
68
69 #Data cleaning - missing values - ignoring the NA values in one example columns
70 nrow(df)
71

```

71:1 (Top Level) ↴

Console Terminal Background Jobs

R 4.3.1 · D:/r/program

```

Standard Deviation of bedrooms: 0.9769223
> #A
> mean_of_review_scores_checkin <- mean(df$review_scores_checkin , na.rm = TRUE)
> median_of_review_scores_checkin <- median(df$review_scores_checkin , na.rm = TRUE)
> sd_of_review_scores_checkin <- sd(df$review_scores_checkin , na.rm = TRUE)
> cat("Mean of bedrooms:", mean_of_review_scores_checkin , "\n")
Mean of bedrooms: 4.813401
> cat("Median of bedrooms:", median_of_review_scores_checkin , "\n")
Median of bedrooms: 4.94
> cat("Standard Deviation of bedrooms:", sd_of_review_scores_checkin , "\n")
Standard Deviation of bedrooms: 0.4125558
> #if all rows with atleast one na value is removed
> df_no_na <- na.omit(df)
> nrow(df_no_na)
[1] 0
> #Data cleaning - missing values - ignoring the NA values in one example columns
> nrow(df)
[1] 39453
>

```

23°C Partly cloudy 22-11-2023

RStudio

File Edt Code View Plots Session Build Debug Profile Tools Help

proj.R*

```

71 colSums(is.na(df))
72 df1 <- subset(df, !(is.na(df$review_scores_checkin)))
73 df1$review_scores_checkin
74 nrow(df1)
75
76 mean_of_bedrooms <- mean(df1$bedrooms , na.rm = TRUE)
77 median_of_bedrooms <- median(df1$bedrooms , na.rm = TRUE)
78 sd_of_bedrooms <- sd(df1$bedrooms , na.rm = TRUE)
79
80 cat("Mean of bedrooms:", mean_of_bedrooms , "\n")
81 cat("Median of bedrooms:", median_of_bedrooms , "\n")
82 cat("Standard Deviation of bedrooms:", sd_of_bedrooms , "\n")
83
84 #Data cleaning - missing values - Filling in a constant value (replacing NA with 0)
85 df_new <- read.csv("listings.csv")
86 df_new$review_scores_checkin <- replace(df$review_scores_checkin, is.na(df$review_scores_checkin), 0)
87
74:1 (Top Level) ↴

```

Console Terminal Background Jobs

R 4.3.1 · D:/r/program

```

calculated_host_listings_count_private_rooms calculated_host_listings_count_shared_rooms
0 0
reviews_per_month
10241
> df1 <- subset(df, !(is.na(df$review_scores_checkin)))
> df1$review_scores_checkin
[1] 4.95 4.95 4.68 4.44 5.00 4.85 4.71 5.00 4.83 5.00 4.75 5.00 5.00 4.67 4.00 4.38 5.00 5.00 4.83
[20] 4.75 4.33 4.75 5.00 4.92 4.97 4.90 4.94 5.00 5.00 4.91 4.96 4.82 4.79 4.94 4.90 4.91 4.88 4.73
[39] 4.45 4.51 4.72 4.79 4.91 4.88 5.00 5.00 5.00 4.82 4.50 5.00 4.75 5.00 4.90 4.88 4.98 4.80 4.93
[58] 5.00 4.86 5.00 5.00 5.00 5.00 4.50 5.00 5.00 4.11 5.00 4.80 5.00 4.78 4.92 4.78 5.00
[77] 5.00 4.86 4.84 4.79 4.94 4.82 4.75 4.96 4.95 4.84 4.79 4.94 4.90 5.00 4.00 4.90 4.29 4.50
[96] 5.00 5.00 5.00 4.89 5.00 4.98 4.97 4.88 5.00 4.86 4.85 4.73 4.71 4.93 5.00 4.40 4.76 4.66 4.79
[115] 4.90 5.00 4.50 5.00 5.00 5.00 4.87 4.89 5.00 4.83 5.00 5.00 4.77 4.92 4.82 4.74 4.73 5.00 5.00
[134] 5.00 5.00 5.00 5.00 4.95 4.83 4.75 5.00 4.93 4.89 4.93 4.87 4.87 4.87 4.80 4.82 4.74 4.94
[153] 4.78 4.86 4.73 4.81 4.67 5.00 4.68 4.50 5.00 4.69 4.92 5.00 5.00 4.95 5.00 4.88 5.00 5.00 4.92
[172] 4.58 4.94 4.85 4.85 4.72 5.00 4.88 4.66 4.60 4.00 5.00 5.00 4.75 4.00 4.86 4.82 4.87 4.83
[191] 4.60 4.88 4.96 4.74 5.00 4.70 4.56 4.67 4.40 5.00 4.60 4.58 4.60 4.64 4.79 5.00 5.00 3.00 5.00
[210] 2.00 5.00 5.00 5.00 4.00 5.00 5.00 5.00 4.00 5.00 5.00 5.00 5.00 5.00 5.00 5.00 5.00 5.00 5.00
[229] 4.00 5.00 5.00 5.00 4.00 5.00 5.00 4.00 5.00 5.00 5.00 4.00 4.00 3.50 5.00 5.00 5.00 5.00 5.00
[248] 5.00 5.00 5.00 5.00 5.00 5.00 5.00 5.00 5.00 5.00 5.00 5.00 5.00 5.00 4.98 5.00 4.73 4.62

```

23°C Partly cloudy 22-11-2023

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

proj.R*

```

75
76 mean_of_bedrooms <- mean(df1$bedrooms, na.rm = TRUE)
77 median_of_bedrooms <- median(df1$bedrooms, na.rm = TRUE)
78 sd_of_bedrooms <- sd(df1$bedrooms, na.rm = TRUE)
79
80 cat("Mean of bedrooms:", mean_of_bedrooms, "\n")
81 cat("Median of bedrooms:", median_of_bedrooms, "\n")
82 cat("Standard Deviation of bedrooms:", sd_of_bedrooms, "\n")
83
84 #Data cleaning - missing values - Filling in a constant value (replacing NA with 0)
85 df_new <- read.csv("listings.csv")
86 df_new$review_scores_checkin <- replace(df$review_scores_checkin, is.na(df$review_scores_checkin),
87 print(df_new$review_scores_checkin)
88 nrow(df_new)
89
90 mean_of_bedrooms <- mean(df_new$bedrooms, na.rm = TRUE)
91
91 (Top Level) ▾

```

Environment History Connections Tutorial

df_no_na 0 obs. of 75 variables

df1 28795 obs. of 75 variables

neighbourhood_fr... 175 obs. of 2 variables

numeric_df 39453 obs. of 37 variables

numeric_df_no_na 16045 obs. of 37 variables

property_type_fr... 78 obs. of 2 variables

room_type_freq_d... 4 obs. of 2 variables

StratifiedSample1 5 obs. of 2 variables

StratifiedSample2 5 obs. of 3 variables

Files Plots Packages Help Viewer Presentation

23°C Partly cloudy 22:57 22-11-2023

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

proj.R*

```

75
76 mean_of_bedrooms <- mean(df1$bedrooms, na.rm = TRUE)
77 median_of_bedrooms <- median(df1$bedrooms, na.rm = TRUE)
78 sd_of_bedrooms <- sd(df1$bedrooms, na.rm = TRUE)
79
80 cat("Mean of bedrooms:", mean_of_bedrooms, "\n")
81 cat("Median of bedrooms:", median_of_bedrooms, "\n")
82 cat("Standard Deviation of bedrooms:", sd_of_bedrooms, "\n")
83
84 #Data cleaning - missing values - Filling in a constant value (replacing NA with 0)
85 df_new <- read.csv("listings.csv")
86 df_new$review_scores_checkin <- replace(df$review_scores_checkin, is.na(df$review_scores_checkin),
87 print(df_new$review_scores_checkin)
88 nrow(df_new)
89
90 mean_of_bedrooms <- mean(df_new$bedrooms, na.rm = TRUE)
91
91 (Top Level) ▾

```

Environment History Connections Tutorial

df_no_na 0 obs. of 75 variables

df1 28795 obs. of 75 variables

neighbourhood_fr... 175 obs. of 2 variables

numeric_df 39453 obs. of 37 variables

numeric_df_no_na 16045 obs. of 37 variables

property_type_fr... 78 obs. of 2 variables

room_type_freq_d... 4 obs. of 2 variables

StratifiedSample1 5 obs. of 2 variables

StratifiedSample2 5 obs. of 3 variables

Files Plots Packages Help Viewer Presentation

23°C Partly cloudy 22:58 22-11-2023

A screenshot of the RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help, and Addins. The left sidebar shows icons for Control Panel, This PC, Network, and TeamViewer. The main area has tabs for proj.R* and Source on Save. The proj.R* tab contains R code for calculating mean, median, and standard deviation of bedrooms and handling missing values. The Environment tab shows global variables like mean_of_host_listings, mean_of_host_total, mean_of_review_scores, median_of_bedrooms, median_of_host_1, median_of_host_2, median_of_host_3, median_of_review, numeric_vars, s, and sd_of_bedrooms. The Console tab shows the execution of the R code, including the output of the print statements and the result of nrow(df_new). The bottom taskbar includes icons for File Explorer, Task View, Start, Search, Edge, Google Chrome, and TeamViewer, along with system status icons for battery, signal, and volume.

A screenshot of the RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help, and Addins. The left sidebar shows icons for User, Control Panel, This PC, Network, Recycle Bin, Microsoft Edge, Google Chrome, and TeamViewer. The main area has tabs for proj.R* and Source on Save. The code editor contains R script code for data cleaning, including reading a CSV file, replacing missing values with 0 or mean, and calculating statistics like mean and standard deviation for the 'bedrooms' column. The Environment browser on the right lists various data frames and their dimensions. The bottom navigation bar includes tabs for Files, Plots, Packages, Help, Viewer, and Presentation, along with a search bar and system status indicators.

The screenshot shows the RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help, and a Project dropdown set to '(None)'. The left sidebar has icons for Control Panel, This PC, Network, and Recycle Bin. The main area has tabs for proj.R* (active), Source on Save, Run, Source, Environment, History, Connections, Tutorial, Global Environment, and a file browser.

Script Editor:

```
83  
84 #Data cleaning - missing values - Filling in a constant value (replacing NA with 0)  
85 df_new <- read.csv("listings.csv")  
86 df_new$review_scores_checkin <- replace(df$review_scores_checkin, is.na(df$review_scores_checkin),  
87 print(df_new$review_scores_checkin)  
88 nrow(df_new)  
89  
90 mean_of_bedrooms <- mean(df_new$bedrooms, na.rm = TRUE)  
91 median_of_bedrooms <- median(df_new$bedrooms, na.rm = TRUE)  
92 sd_of_bedrooms <- sd(df_new$bedrooms, na.rm = TRUE)  
93  
94 cat("Mean of bedrooms:", mean_of_bedrooms, "\n")  
95 cat("Median of bedrooms:", median_of_bedrooms, "\n")  
96 cat("Standard Deviation of bedrooms:", sd_of_bedrooms, "\n")  
97  
98 #Data cleaning - missing values - Filling in mean value (replacing NA with mean)  
99
```

Environment Browser:

Object	Type	Size
df_new	data frame	39453 obs. of 75 variables
df_new1	data frame	39453 obs. of 75 variables
df_no_duplicates...	data frame	39453 obs. of 75 variables
df_no_duplicates...	data frame	39453 obs. of 75 variables
df_no_na	data frame	0 obs. of 75 variables
df1	data frame	28795 obs. of 75 variables
neighbourhood_fr...	data frame	175 obs. of 2 variables
numeric_df	data frame	39453 obs. of 37 variables
numeric_df_no_na	data frame	16045 obs. of 37 variables

Console:

```
[590] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00  
[609] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00  
[628] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00  
[647] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00  
[666] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00  
[685] 0.00 5.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00  
[704] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00  
[723] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00  
[742] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00  
[761] 4.73 4.62 4.66 4.75 4.75 0.00 0.00 4.93 0.00 5.00 4.69 4.59 4.80 4.29 4.25 5.00 5.00 4.71 4.25  
[780] 4.88 4.86 4.67 4.57 4.67 4.60 4.78 5.00 4.81 4.75 4.80 5.00 4.75 0.00 4.33 0.00 4.77 4.68 4.74  
[799] 0.00 5.00 4.76 4.44 4.47 4.70 4.62 4.64 4.59 4.50 4.83 4.51 4.86 4.78 4.59 4.25 4.68 4.78 4.75  
[818] 0.00 4.70 4.83 4.51 4.33 4.78 4.68 4.86 4.65 4.70 4.73 4.61 4.77 4.71 4.43 4.57 4.58 4.90 4.44  
[837] 4.69 4.78 4.91 4.88 4.67 5.00 4.63 4.64 4.83 0.00 5.00 4.95 5.00 4.86 4.97 4.85 5.00 4.98 4.75  
[856] 4.90 4.83 4.95 4.96 4.86 4.95 4.81 4.91 4.90 4.91 4.91 4.89 4.88 4.99 4.98 5.00 5.00 4.93 5.00  
[875] 4.92 4.98 0.00 0.00 0.00 4.91 0.00 0.00 4.56 5.00 4.91 5.00 5.00 4.98 4.80 4.92 4.88 5.00 4.54  
[894] 5.00 0.00 4.97 5.00 4.63 5.00 0.00 5.00 0.00 4.60 0.00 4.85 4.62 4.85 4.65 4.55 4.59 4.64 4.68  
[913] 4.80 4.89 4.73 4.89 4.80 5.00 4.76 4.79 4.84 4.98 5.00 4.78 4.97 5.00 5.00 4.94 4.93 4.92 4.70  
[932] 4.82 4.86 5.00 4.82 4.76 4.68 4.69 4.92 4.95 4.99 4.93 0.00 3.87 4.94 4.93 4.96 4.97 4.58 4.91  
[951] 4.93 0.00 4.77 4.87 4.74 4.96 4.94 4.97 4.93 5.00 4.95 5.00 5.00 4.91 5.00 5.00 4.97 4.99
```

The screenshot shows an RStudio interface with the following details:

- File Menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Toolbar:** Includes icons for file operations like Open, Save, Print, and Addins.
- Code Editor:** A tab labeled "proj.R*" containing R code for data cleaning and printing descriptive statistics.
- Environment Tab:** Shows the global environment with various objects and their characteristics.
- Console Tab:** Displays the R session output, including the results of the printed objects.
- System Status Bar:** Shows the date (22-11-2023), time (23:01), battery level (23%), and network status (Partly cloudy).

The screenshot shows an RStudio interface with the following details:

- File Menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Toolbar:** Includes icons for file operations like Open, Save, Print, and Addins.
- Code Editor:** A script editor titled "proj.R" containing R code for calculating mean, median, and standard deviation of bedrooms, and handling missing values.
- Environment View:** Shows the global environment with various objects and their characteristics.
- Console View:** Displays the R command-line interface with the R version (R 4.3.1), current working directory (D:\r\prog\r), and the results of the executed R code.
- System Tray:** Shows the date (22-11-2023), time (23:01), battery status (ENG IN), signal strength, and a small icon for the application.

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

proj.R*

```

100 df_new1$review_scores_checkin <- replace(df_new1$review_scores_checkin, is.na(df_new1$review_scores_checkin), 4)
101 print(df_new1$review_scores_checkin)
102 nrow(df_new1)
103
104 mean_of_bedrooms <- mean(df_new1$bedrooms, na.rm = TRUE)
105 median_of_bedrooms <- median(df_new1$bedrooms, na.rm = TRUE)
106 sd_of_bedrooms <- sd(df_new1$bedrooms, na.rm = TRUE)
107
108 cat("Mean of bedrooms:", mean_of_bedrooms, "\n")
109 cat("Median of bedrooms:", median_of_bedrooms, "\n")
110 cat("Standard Deviation of bedrooms:", sd_of_bedrooms, "\n")
111
112
113 #Stratified sampling: population is split into groups and a certain number of members from each group
#selected to be included in the sample.
114 library(sampling)
115
116
102:1 (Top Level) R Script
```

Console Terminal Background Jobs

R 4.3.1 - D:/r/prog/

```

> print(df_new1$review_scores_checkin)
[1] 4.950000 4.813401 4.950000 4.680000 4.440000 5.000000 4.850000 4.813401 4.710000 5.000000
[11] 4.813401 4.830000 5.000000 4.813401 4.950000 5.000000 4.670000 4.000000 4.380000
[21] 5.000000 5.000000 4.830000 4.750000 4.330000 4.750000 5.000000 4.920000 4.970000 4.813401
[31] 4.900000 4.813401 4.813401 4.940000 5.000000 5.000000 4.910000 4.960000 4.820000 4.790000
[41] 4.940000 4.900000 4.910000 4.880000 4.730000 4.450000 4.813401 4.510000 4.720000 4.790000
[51] 4.910000 4.880000 5.000000 4.813401 5.000000 4.820000 4.500000 5.000000 4.750000
[61] 5.000000 4.900000 4.880000 4.980000 4.800000 4.930000 4.813401 5.000000 4.860000 4.813401
[71] 5.000000 5.000000 5.000000 5.000000 5.000000 4.813401 4.500000 5.000000 5.000000
[81] 4.110000 5.000000 4.813401 4.813401 4.800000 5.000000 4.780000 4.920000 4.780000
[91] 4.813401 4.813401 4.813401 5.000000 5.000000 4.860000 4.840000 4.790000 4.940000 4.820000
[101] 4.750000 4.960000 4.950000 4.840000 4.790000 4.940000 4.940000 4.970000 5.000000 4.000000
[111] 4.900000 4.290000 4.500000 5.000000 5.000000 5.000000 4.890000 5.000000 4.813401 4.980000
[121] 4.970000 4.880000 5.000000 4.860000 4.850000 4.730000 4.813401 4.813401 4.710000 4.950000
[131] 5.000000 4.400000 4.760000 4.660000 4.790000 4.900000 4.813401 4.813401 5.000000 4.500000
[141] 5.000000 5.000000 4.870000 4.890000 5.000000 4.830000 5.000000 5.000000 4.770000
[151] 4.920000 4.820000 4.740000 4.730000 5.000000 5.000000 5.000000 5.000000 5.000000
[161] 5.000000 4.950000 4.830000 4.750000 5.000000 4.930000 4.890000 4.930000 4.870000 4.813401
[171] 4.870000 4.870000 4.800000 4.820000 4.740000 4.940000 4.780000 4.860000 4.730000 4.810000
[181] 4.670000 5.000000 4.680000 4.500000 5.000000 4.813401 4.813401 4.690000 4.920000 5.000000

```

23°C Partly cloudy 23:02 22-11-2023 ENG IN

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

proj.R*

```

100 df_new1$review_scores_checkin <- replace(df_new1$review_scores_checkin, is.na(df_new1$review_scores_checkin), 4)
101 print(df_new1$review_scores_checkin)
102 nrow(df_new1)
103
104 mean_of_bedrooms <- mean(df_new1$bedrooms, na.rm = TRUE)
105 median_of_bedrooms <- median(df_new1$bedrooms, na.rm = TRUE)
106 sd_of_bedrooms <- sd(df_new1$bedrooms, na.rm = TRUE)
107
108 cat("Mean of bedrooms:", mean_of_bedrooms, "\n")
109 cat("Median of bedrooms:", median_of_bedrooms, "\n")
110 cat("Standard Deviation of bedrooms:", sd_of_bedrooms, "\n")
111
112
113 #Stratified sampling: population is split into groups and a certain number of members from each group
#selected to be included in the sample.
114 library(sampling)
115
116
102:1 (Top Level) R Script
```

Console Terminal Background Jobs

R 4.3.1 - D:/r/prog/

```

> print(df_new1$review_scores_checkin)
[1] 4.670000 5.000000 4.680000 4.500000 5.000000 4.813401 4.813401 4.690000 4.920000 5.000000
[191] 5.000000 4.813401 4.950000 4.813401 5.000000 4.880000 5.000000 4.920000 4.560000
[201] 4.940000 4.850000 4.850000 4.720000 5.000000 4.880000 4.660000 4.680000 4.000000 5.000000
[211] 5.000000 4.813401 5.000000 4.750000 4.813401 4.000000 4.813401 4.813401 4.860000 4.860000
[221] 4.870000 4.730000 4.600000 4.880000 4.790000 4.740000 5.000000 4.700000 4.560000 4.670000
[231] 4.400000 5.000000 4.813401 4.600000 4.580000 4.600000 4.640000 4.790000 4.813401 4.813401
[241] 4.813401 4.813401 5.000000 4.813401 4.813401 4.813401 4.813401 4.813401 4.813401 4.813401
[251] 4.813401 4.813401 4.813401 4.813401 4.813401 4.813401 4.813401 4.813401 4.813401 4.813401
[261] 4.813401 4.813401 4.813401 5.000000 4.813401 4.813401 4.813401 4.813401 4.813401 4.813401
[271] 4.813401 4.813401 3.000000 4.813401 5.000000 4.813401 4.813401 2.000000 4.813401 5.000000
[281] 4.813401 5.000000 4.813401 5.000000 4.813401 4.813401 4.813401 4.813401 5.000000 4.813401
[291] 4.813401 4.813401 4.813401 4.813401 4.813401 5.000000 4.813401 4.813401 4.813401 5.000000
[301] 5.000000 4.813401 4.000000 4.813401 4.813401 4.813401 4.813401 4.813401 4.813401 4.813401
[311] 4.813401 4.813401 4.813401 4.813401 4.813401 4.813401 4.813401 4.813401 4.813401 4.813401
[321] 4.813401 4.813401 4.813401 4.813401 5.000000 4.813401 4.813401 4.813401 4.813401 4.813401
[331] 4.813401 4.813401 4.813401 4.813401 4.813401 4.813401 4.813401 4.813401 4.813401 4.813401
[341] 4.813401 4.813401 4.813401 5.000000 4.813401 4.813401 4.813401 4.813401 4.813401 4.813401
[351] 5.000000 4.813401 4.813401 5.000000 4.813401 5.000000 4.813401 5.000000 5.000000 5.000000
[361] 5.000000 4.000000 4.813401 5.000000 4.813401 4.813401 4.813401 5.000000 4.813401 4.813401

```

23°C Partly cloudy 23:03 22-11-2023 ENG IN

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source Save Run Source Addins
proj.R*
100 df_new1$review_scores_checkin <- replace(df_new1$review_scores_checkin, is.na(df_new1$review_scores))
101 print(df_new1$review_scores_checkin)
102 nrow(df_new1)
103
104 mean_of_bedrooms <- mean(df_new1$bedrooms, na.rm = TRUE)
105 median_of_bedrooms <- median(df_new1$bedrooms, na.rm = TRUE)
106 sd_of_bedrooms <- sd(df_new1$bedrooms, na.rm = TRUE)
107
108 cat("Mean of bedrooms:", mean_of_bedrooms, "\n")
109 cat("Median of bedrooms:", median_of_bedrooms, "\n")
110 cat("Standard Deviation of bedrooms:", sd_of_bedrooms, "\n")
111
112
113 #Stratified sampling: population is split into groups and a certain number of members from each group
#selected to be included in the sample.
114 library(sampling)
115
116 <
102:1 (Top Level) R Script

```

Environment History Connections Tutorial

df_new1 39453 obs. of 75 variables
df_no_duplicates... 39453 obs. of 75 variables
df_no_duplicates... 39453 obs. of 75 variables
df_no_na 0 obs. of 75 variables
df1 28795 obs. of 75 variables
neighbourhood_fr... 175 obs. of 2 variables
numeric_df 39453 obs. of 37 variables
numeric_df_no_na 16045 obs. of 37 variables
property_type_fr... 78 obs. of 2 variables

Files Plots Packages Help Viewer Presentation

23°C Partly cloudy 23:03 22-11-2023 ENG IN

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source Save Run Source Addins
proj.R*
100 df_new1$review_scores_checkin <- replace(df_new1$review_scores_checkin, is.na(df_new1$review_scores))
101 print(df_new1$review_scores_checkin)
102 nrow(df_new1)
103
104 mean_of_bedrooms <- mean(df_new1$bedrooms, na.rm = TRUE)
105 median_of_bedrooms <- median(df_new1$bedrooms, na.rm = TRUE)
106 sd_of_bedrooms <- sd(df_new1$bedrooms, na.rm = TRUE)
107
108 cat("Mean of bedrooms:", mean_of_bedrooms, "\n")
109 cat("Median of bedrooms:", median_of_bedrooms, "\n")
110 cat("Standard Deviation of bedrooms:", sd_of_bedrooms, "\n")
111
112
113 #Stratified sampling: population is split into groups and a certain number of members from each group
#selected to be included in the sample.
114 library(sampling)
115
116 <
102:1 (Top Level) R Script

```

Environment History Connections Tutorial

df_new1 39453 obs. of 75 variables
df_no_duplicates... 39453 obs. of 75 variables
df_no_duplicates... 39453 obs. of 75 variables
df_no_na 0 obs. of 75 variables
df1 28795 obs. of 75 variables
neighbourhood_fr... 175 obs. of 2 variables
numeric_df 39453 obs. of 37 variables
numeric_df_no_na 16045 obs. of 37 variables
property_type_fr... 78 obs. of 2 variables

Files Plots Packages Help Viewer Presentation

23°C Partly cloudy 23:04 22-11-2023 ENG IN

The screenshot shows the RStudio interface with the following details:

- Title Bar:** R Studio
- File Menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help
- Toolbar:** Source on Save, Run, Source
- Code Editor (proj.R*):**

```
104 mean_of_bedrooms <- mean(df_new1$bedrooms, na.rm = TRUE)
105 median_of_bedrooms <- median(df_new1$bedrooms, na.rm = TRUE)
106 sd_of_bedrooms <- sd(df_new1$bedrooms, na.rm = TRUE)
107
108 cat("Mean of bedrooms:", mean_of_bedrooms, "\n")
109 cat("Median of bedrooms:", median_of_bedrooms, "\n")
110 cat("Standard Deviation of bedrooms:", sd_of_bedrooms, "\n")
111
112
113 #Stratified sampling: population is split into groups and a certain number of members from each group
114 #selected to be included in the sample.
115 library(sampling)
116 stratifiedSample1 <- strata(data = df, size = 5, method = "srswor")
117 df[StratifiedSample1$ID_unit, ]
118
119 StratifiedSample2 <- strata(data = df, size = 5, method = "srswr")
120
```
- Environment Pane:** Shows variables like mean_of_host_list, mean_of_host_tot, mean_of_review_s, median_of_bedrooms, median_of_host_1, median_of_host_2, median_of_host_3, median_of_review_4, numeric_vars, s, and sd_of_bedrooms.
- Console Pane:** Shows the execution of R code and its output. The output includes the mean, median, and standard deviation of bedrooms, and the results of stratified sampling.
- System Tray:** Shows the date (23/05/22), weather (Partly cloudy), and system status (ENG IN IN).

5 random rows are printed all screenshots of row data printed is not included in both sample 1 and 2

The screenshot shows the RStudio interface with the following details:

- Top Bar:** R Studio, File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Environment Tab:** Shows the Global Environment with objects like `df_no_na`, `df1`, `neighbourhood_fr...`, `numeric_df`, etc.
- Code Editor:** Displays R script code for stratified sampling and data imputation.
- Console Tab:** Shows the output of the R script, including the mean calculation for each column and the final stratified sample.
- Data View:** Shows the `df` data frame with columns: `id`, `listing_url`, `scrape_id`, `last_scraped`, `source`, and `name`.

The screenshot shows an RStudio interface with the following details:

- File Menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Toolbar:** Source on Save, Run, Source, Addins.
- Code Editor:** A script named "proj.R" containing R code for stratified sampling. The code includes library(sampling), StratifiedSample1, StratifiedSample2, and a cor() matrix calculation.
- Environment Tab:** Shows the Global Environment with objects like OT_no_na, df1, neighbourhood_fr..., numeric_df, numeric_df_no_na, property_type_fr..., room_type_freq_d, StratifiedSample1, and StratifiedSample2.
- Console Tab:** Displays the output of the R code, including the cor() matrix and a summary of the data frame structure.
- Plots Tab:** Contains a scatter plot of host_total_listings_count vs bedrooms.
- Packages Tab:** Lists available packages: tidyverse, dplyr, purrr, readr, tibble, ggplot2, tidyverse, dplyr, purrr, readr, tibble, and rlang.
- Help Tab:** Provides help for the cor() function.
- Viewer Tab:** Shows the correlation matrix output.
- Presentation Tab:** Provides presentation tools.

The screenshot shows the RStudio interface with the following components:

- File Menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Toolbar:** Source on Save, Run, Source, Environment, History, Connections, Tutorial, Project: (None).
- Script Editor:** A code editor window titled "proj.R" containing R code for imputing missing values and calculating correlation matrices.
- Console:** Displays the execution of the R code from the script editor.
- Environment Browser:** A table showing the global environment with objects like df_imputed, cor_matrix_imputed, and df1.
- Bottom Status Bar:** Shows system information including the date (22-11-2023), time (23:10), and weather (23°C Partly cloudy).

RStudio Environment View:

```

proj.R*:
134 # Example 1: Frequency distribution for neighbourhood
135 neighbourhood_freq_dplyr <- df %>% count(neighbourhood)
136 neighbourhood_freq_dplyr
137 neighbourhood_freq_dplyr
138
139 # Example 2: Frequency distribution for property_type
140 property_type_freq_dplyr <- df %>% count(property_type)
141 property_type_freq_dplyr
142
143 <
139:1 (Top Level) :

```

RStudio Console View:

```

R 4.3.1 - D:/r/prog/
100 0.016829160 1.00000000 0.02682/21
review_scores_rating 0.012934306 0.02682721 1.00000000
> # Example 1: Frequency distribution for neighbourhood
> neighbourhood_freq_dplyr <- df %>% count(neighbourhood)
> neighbourhood_freq_dplyr

```

RStudio Environment View (Continued):

neighbourhood	n
Bronx, New York, United States	1
Bushwick, Brooklyn, New York, United States	1
Crown Heights, NY, New York, United States	1
Elmhurst, New York, United States	1
North Queens, New York, United States	1
Springfield Gardens, New York, United States	1
8425 Elmhurst avenue, New York, United States	1
ASTORIA, New York, United States	1
Arverne, New York, United States	3
Astoria, New York, United States	13
Astoria, Queens, New York, United States	1
Astoria Queens, New York, United States	1
Astoria, New York, United States	51
Astoria, Queens, New York, United States	2
BROOKLYN, New York, United States	2
BROOKLYN, New York, United States	1
Bayside, New York, United States	1
Briarwood, New York, United States	2
Brooklyn, New York, United States	1
Bronx, New York, United States	15

all 175 rows not included in screenshot

RStudio Environment View:

```

proj.R*:
138
139 # Example 2: Frequency distribution for property_type
140 property_type_freq_dplyr <- df %>% count(property_type)
141 property_type_freq_dplyr
142
143 # Example 3: Frequency distribution for room_type
144 room_type_freq_dplyr <- df %>% count(room_type)
145 room_type_freq_dplyr
146
147 <
143:1 (Top Level) :

```

RStudio Console View:

```

R 4.3.1 - D:/r/prog/
165 astoria, New York, United States 2
166 bronx, New York, United States 1
167 bronx, New York, United States 3
168 brooklyn, New York, United States 10
169 elmhurst, Queens, New York, United States 1
170 flushing, New York, United States 4
171 hollis, New York, United States 1
172 queens, New York, United States 1
173 woodside, New York, United States 1
174 纽约, New York, United States 2
175 纽约市, New York, United States 1
> # Example 2: Frequency distribution for property_type
> property_type_freq_dplyr <- df %>% count(property_type)
> property_type_freq_dplyr

```

RStudio Environment View (Continued):

property_type	n
Barn	2
Boat	6
Camper/RV	11
Casa particular	6
Castle	1
Cave	1
Entire bed and breakfast	1
Entire bungalow	15
Entire condo	1323
Entire cottage	7
Entire guest suite	340

all 78 rows not included in screenshots

Visual Analysis

The screenshot shows the RStudio interface with the following components:

- File Explorer (Left):** Shows icons for Control Panel, This PC, Network, and Recycle Bin.
- Code Editor (Top Left):** Displays R script code for visual analysis, including histograms for numerical variables. The code uses `par(mfrow = c(2,2))` and `hist` functions with lightblue fills and black borders.
- Console (Bottom Left):** Shows the R command history, including the creation of a frequency distribution for neighbourhood_group_cleansed and the execution of the histogram code.
- Environment (Top Right):** Shows the global environment with objects like `local_type_ideq_u` (4 obs. of 2 variables), `StratifiedSample1` (5 obs. of 2 variables), and `StratifiedSample2` (5 obs. of 3 variables). It also lists values for `cross_tab1` through `cross_tab4`.
- Plots (Bottom Right):** Four histograms are displayed:
 - id:** Frequency distribution of `id` values, ranging from 0e+00 to 8e+17.
 - scrape_id:** Frequency distribution of `scrape_id` values, ranging from 2.00e+13 to 2.20e+13.
 - host_id:** Frequency distribution of `host_id` values, ranging from 0e+00 to 4e+08.
 - host_listings_count:** Frequency distribution of `host_listings_count` values, ranging from 0 to 3000.

The screenshot shows the RStudio interface with the following details:

- File Menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Addins:** Source on Save, Supply, Run, Source.
- Code Editor:** A script named "proj.R" containing R code for histogram generation. The code uses `grid` and `gridExtra` packages to create a grid of histograms for numerical variables.
- Environment:** Shows the global environment with objects like `cross_tab1`, `cross_tab2`, `cross_tab3`, and `cross_tab4`.
- Console:** Displays the output of the R code, showing histograms for variables like `host_total_listings_count`, `latitude`, `longitude`, and `accommodates`.
- Plots:** Four histograms are displayed in the main pane:
 - host_total_listings_count:** Frequency vs. host_total_listings_count (0 to 30,000).
 - latitude:** Frequency vs. latitude (40.5 to 40.9).
 - longitude:** Frequency vs. longitude (-74.3 to -73.7).
 - accommodates:** Frequency vs. accommodates (0 to 15).
- System Status:** 23°C, Partly cloudy; ENG IN; 22-11-2023; 23:15.

RStudio

```

proj.R* x
186+ for (i in 5:8) {
187+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
188+ }
189 par(mfrow = c(2, 2))
190+ for (i in 9:12) {
191+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
192+ }
193 par(mfrow = c(2, 2))
194+ for (i in 13:16) {
195+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
196+ }
197 par(mfrow = c(2, 2))
198+ for (i in 17:20) {
199+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
200+ }
201+ 
```

197:1 (Top Level) ↴

Console Terminal Background Jobs

```

R 4.3.1 - D:/r/prog/
> numeric_vars <- sapply(df, is.numeric)
> numeric_df <- df[, numeric_vars]
# Histograms
> par(mfrow = c(2,2))
> for (i in 1:4) {
+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
+ }
> par(mfrow = c(2,2))
> for (i in 5:8) {
+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
+ }
> par(mfrow = c(2,2))
> for (i in 9:12) {
+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
+ }
> par(mfrow = c(2,2))
> for (i in 13:16) {
+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
+ }
> par(mfrow = c(2,2))
> for (i in 17:20) {
+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
+ }
> 
```

16L

Environment History Connections Tutorial

Global Environment

- room_type_1eq_u_&_os_ 5 obs. of 2 variables
- StratifiedSample1 5 obs. of 2 variables
- StratifiedSample2 5 obs. of 3 variables

Values

col	review_scores_rating
cross_tab1	"table" int [1:5, 1:4] 628 7563 10933 2623 23...
cross_tab2	"table" int [1:3, 1:2] 420 24635 6556 77 6141...
cross_tab3	"table" int [1:175, 1:3] 162 0 0 0 0 0 0 0 0 ...
cross_tab4	"table" int [1:5, 1:4] 628 7563 10933 2623 23...

Files Plots Packages Help Viewer Presentation

minimum_minimum_nights maximum_minimum_nights

minimum_maximum_nights maximum_maximum_nights

Frequency

Frequency

Frequency

Frequency

23°C Partly cloudy 22-11-2023 23:16 ENG IN

RStudio

```

proj.R* x
190+ for (i in 9:12) {
191+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
192+ }
193 par(mfrow = c(2, 2))
194+ for (i in 13:16) {
195+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
196+ }
197 par(mfrow = c(2, 2))
198+ for (i in 17:20) {
199+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
200+ }
201 par(mfrow = c(2, 2))
202+ for (i in 21:24) {
203+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
204+ }
205+ 
```

201:1 (Top Level) ↴

Console Terminal Background Jobs

```

R 4.3.1 - D:/r/prog/
> for (i in 1:4) {
+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
+ }
> par(mfrow = c(2, 2))
> for (i in 5:8) {
+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
+ }
> par(mfrow = c(2, 2))
> for (i in 9:12) {
+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
+ }
> par(mfrow = c(2, 2))
> for (i in 13:16) {
+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
+ }
> par(mfrow = c(2, 2))
> for (i in 17:20) {
+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
+ }
> 
```

20L

Environment History Connections Tutorial

Global Environment

- room_type_1eq_u_&_os_ 5 obs. of 2 variables
- StratifiedSample1 5 obs. of 2 variables
- StratifiedSample2 5 obs. of 3 variables

Values

col	review_scores_rating
cross_tab1	"table" int [1:5, 1:4] 628 7563 10933 2623 23...
cross_tab2	"table" int [1:3, 1:2] 420 24635 6556 77 6141...
cross_tab3	"table" int [1:175, 1:3] 162 0 0 0 0 0 0 0 0 ...
cross_tab4	"table" int [1:5, 1:4] 628 7563 10933 2623 23...

Files Plots Packages Help Viewer Presentation

minimum_nights_avg_ntm maximum_nights_avg_ntm

availability_30 availability_60

Frequency

Frequency

Frequency

Frequency

23°C Partly cloudy 22-11-2023 23:17 ENG IN

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

projR.R

```

194+ for (i in 13:16) {
195+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
196+ }
197 par(mfrow = c(2, 2))
198+ for (i in 17:20) {
199+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
200+ }
201 par(mfrow = c(2, 2))
202+ for (i in 21:24) {
203+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
204+ }
205 par(mfrow = c(2, 2))
206+ for (i in 25:28) {
207+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
208+ }
209+
209:1 (Top Level) ◊

```

Console Terminal Background Jobs

R 4.3.1 - D:/r/prop/ ◊

```

> for (i in 5:8) {
+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
+ }
+ par(mfrow = c(2, 2))
+ for (i in 9:12) {
+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
+ }
+ par(mfrow = c(2, 2))
+ for (i in 13:16) {
+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
+ }
+ par(mfrow = c(2, 2))
+ for (i in 17:20) {
+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
+ }
+ par(mfrow = c(2, 2))
+ for (i in 21:24) {
+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
+ }
+ par(mfrow = c(2, 2))
+ for (i in 25:28) {
+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
+ }
> |

```

Environment History Connections Tutorial

Global Environment

- review_scores_rating
- StratifiedSample1
- StratifiedSample2

Values

col	review_scores_rating
cross_tab1	table [1:5, 1:4] 628 7563 10933 2623 23...
cross_tab2	table [1:3, 1:2] 420 24635 6556 77 6141...
cross_tab3	table [1:175, 1:3] 162 0 0 0 0 0 0 0 0 ...
cross_tab4	table [1:5, 1:4] 628 7563 10933 2623 23...

Files Plots Packages Help Viewer Presentation

availability_90 availability_365

number_of_reviews number_of_reviews_ltm

23°C Partly cloudy ENG IN 23:17 22-11-2023

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

projR.R

```

198+ for (i in 17:20) {
199+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
200+ }
201 par(mfrow = c(2, 2))
202+ for (i in 21:24) {
203+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
204+ }
205 par(mfrow = c(2, 2))
206+ for (i in 25:28) {
207+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
208+ }
209 par(mfrow = c(2, 2))
210+ for (i in 29:32) {
211+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
212+ }
213+
209:1 (Top Level) ◊

```

Console Terminal Background Jobs

R 4.3.1 - D:/r/prop/ ◊

```

> for (i in 9:12) {
+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
+ }
+ par(mfrow = c(2, 2))
+ for (i in 13:16) {
+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
+ }
+ par(mfrow = c(2, 2))
+ for (i in 17:20) {
+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
+ }
+ par(mfrow = c(2, 2))
+ for (i in 21:24) {
+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
+ }
+ par(mfrow = c(2, 2))
+ for (i in 25:28) {
+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
+ }
> |

```

Environment History Connections Tutorial

Global Environment

- review_scores_rating
- StratifiedSample1
- StratifiedSample2

Values

col	review_scores_rating
cross_tab1	table [1:5, 1:4] 628 7563 10933 2623 23...
cross_tab2	table [1:3, 1:2] 420 24635 6556 77 6141...
cross_tab3	table [1:175, 1:3] 162 0 0 0 0 0 0 0 0 ...
cross_tab4	table [1:5, 1:4] 628 7563 10933 2623 23...

Files Plots Packages Help Viewer Presentation

number_of_reviews_l30d review_scores_rating

review_scores_accuracy review_scores_cleanliness

23°C Partly cloudy ENG IN 23:18 22-11-2023

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

proj.R

```

202+ for (i in 21:24) {
203+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
204+
205+ par(mfrow = c(2, 2))
206+ for (i in 25:28) {
207+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
208+ }
209+ par(mfrow = c(2, 2))
210+ for (i in 29:32) {
211+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
212+ }
213+ par(mfrow = c(2, 2))
214+ for (i in 33:36) {
215+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
216+ }
217+
218+ # Density plots
219+ # Identify numeric variables
220+ numeric_vars <- sapply(df, is.numeric)
221+ numeric_df <- df[, numeric_vars]
222+
223+ par(mfrow = c(2, 2))
224+ for (i in 1:20) {
225+   density(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
226+ }
227+ par(mfrow = c(2, 2))
228+ for (i in 21:24) {
229+   density(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
230+ }
231+ par(mfrow = c(2, 2))
232+ for (i in 25:28) {
233+   density(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
234+ }
235+ par(mfrow = c(2, 2))
236+ for (i in 29:32) {
237+   density(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
238+ }
239+ par(mfrow = c(2, 2))
240+ for (i in 33:36) {
241+   density(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
242+ }
243+
244+ # Histograms
245+ for (i in 1:20) {
246+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
247+ }
248+ par(mfrow = c(2, 2))
249+ for (i in 21:24) {
250+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
251+ }
252+ par(mfrow = c(2, 2))
253+ for (i in 25:28) {
254+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
255+ }
256+ par(mfrow = c(2, 2))
257+ for (i in 29:32) {
258+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
259+ }
260+ par(mfrow = c(2, 2))
261+ for (i in 33:36) {
262+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
263+ }
264+
265+ # Boxplots
266+ for (i in 1:20) {
267+   boxplot(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
268+ }
269+ par(mfrow = c(2, 2))
270+ for (i in 21:24) {
271+   boxplot(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
272+ }
273+ par(mfrow = c(2, 2))
274+ for (i in 25:28) {
275+   boxplot(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
276+ }
277+ par(mfrow = c(2, 2))
278+ for (i in 29:32) {
279+   boxplot(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
280+ }
281+ par(mfrow = c(2, 2))
282+ for (i in 33:36) {
283+   boxplot(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
284+ }
285+
286+ # Scatter plots
287+ for (i in 1:20) {
288+   scatterplot(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
289+ }
290+ par(mfrow = c(2, 2))
291+ for (i in 21:24) {
292+   scatterplot(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
293+ }
294+ par(mfrow = c(2, 2))
295+ for (i in 25:28) {
296+   scatterplot(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
297+ }
298+ par(mfrow = c(2, 2))
299+ for (i in 29:32) {
300+   scatterplot(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
301+ }
302+ par(mfrow = c(2, 2))
303+ for (i in 33:36) {
304+   scatterplot(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
305+ }
306+
307+ # Correlation matrix
308+ cor_matrix <- cor(numeric_df)
309+ cor_matrix
310+
311+ # Heatmap
312+ heatmap(cor_matrix, col = "lightblue", border = "black")
313+
314+ # Summary statistics
315+ summary(numeric_df)
316+
317+ # Descriptive statistics
318+ describe(numeric_df)
319+
320+ # Data visualization
321+ plot(numeric_df)
322+
323+ # Data analysis
324+ analyze(numeric_df)
325+
326+ # Data mining
327+ mine(numeric_df)
328+
329+ # Data science
330+ science(numeric_df)
331+
332+ # Data engineering
333+ engineer(numeric_df)
334+
335+ # Data modeling
336+ model(numeric_df)
337+
338+ # Data prediction
339+ predict(numeric_df)
340+
341+ # Data inference
342+ infer(numeric_df)
343+
344+ # Data visualization
345+ viz(numeric_df)
346+
347+ # Data analysis
348+ anal(numeric_df)
349+
350+ # Data mining
351+ min(numeric_df)
352+
353+ # Data science
354+ sci(numeric_df)
355+
356+ # Data engineering
357+ eng(numeric_df)
358+
359+ # Data modeling
360+ mod(numeric_df)
361+
362+ # Data prediction
363+ pred(numeric_df)
364+
365+ # Data inference
366+ inf(numeric_df)
367+
368+ # Data visualization
369+ vis(numeric_df)
370+
371+ # Data analysis
372+ anal(numeric_df)
373+
374+ # Data mining
375+ min(numeric_df)
376+
377+ # Data science
378+ sci(numeric_df)
379+
380+ # Data engineering
381+ eng(numeric_df)
382+
383+ # Data modeling
384+ mod(numeric_df)
385+
386+ # Data prediction
387+ pred(numeric_df)
388+
389+ # Data inference
390+ inf(numeric_df)
391+
392+ # Data visualization
393+ vis(numeric_df)
394+
395+ # Data analysis
396+ anal(numeric_df)
397+
398+ # Data mining
399+ min(numeric_df)
400+
401+ # Data science
402+ sci(numeric_df)
403+
404+ # Data engineering
405+ eng(numeric_df)
406+
407+ # Data modeling
408+ mod(numeric_df)
409+
410+ # Data prediction
411+ pred(numeric_df)
412+
413+ # Data inference
414+ inf(numeric_df)
415+
416+ # Data visualization
417+ vis(numeric_df)
418+
419+ # Data analysis
420+ anal(numeric_df)
421+
422+ # Data mining
423+ min(numeric_df)
424+
425+ # Data science
426+ sci(numeric_df)
427+
428+ # Data engineering
429+ eng(numeric_df)
430+
431+ # Data modeling
432+ mod(numeric_df)
433+
434+ # Data prediction
435+ pred(numeric_df)
436+
437+ # Data inference
438+ inf(numeric_df)
439+
440+ # Data visualization
441+ vis(numeric_df)
442+
443+ # Data analysis
444+ anal(numeric_df)
445+
446+ # Data mining
447+ min(numeric_df)
448+
449+ # Data science
450+ sci(numeric_df)
451+
452+ # Data engineering
453+ eng(numeric_df)
454+
455+ # Data modeling
456+ mod(numeric_df)
457+
458+ # Data prediction
459+ pred(numeric_df)
460+
461+ # Data inference
462+ inf(numeric_df)
463+
464+ # Data visualization
465+ vis(numeric_df)
466+
467+ # Data analysis
468+ anal(numeric_df)
469+
470+ # Data mining
471+ min(numeric_df)
472+
473+ # Data science
474+ sci(numeric_df)
475+
476+ # Data engineering
477+ eng(numeric_df)
478+
479+ # Data modeling
480+ mod(numeric_df)
481+
482+ # Data prediction
483+ pred(numeric_df)
484+
485+ # Data inference
486+ inf(numeric_df)
487+
488+ # Data visualization
489+ vis(numeric_df)
490+
491+ # Data analysis
492+ anal(numeric_df)
493+
494+ # Data mining
495+ min(numeric_df)
496+
497+ # Data science
498+ sci(numeric_df)
499+
499+ (Top Level) ◊

```

Environment History Connections Tutorial

review_scores_rating

review_scores_checkin

review_scores_communication

review_scores_location

review_scores_value

Files Plots Packages Help Viewer Presentation

ENG IN 22-11-2023

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

proj.R

```

202+ for (i in 1:20) {
203+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
204+
205+ par(mfrow = c(2, 2))
206+ for (i in 21:24) {
207+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
208+ }
209+ par(mfrow = c(2, 2))
210+ for (i in 25:28) {
211+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
212+ }
213+ par(mfrow = c(2, 2))
214+ for (i in 29:32) {
215+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
216+ }
217+
218+ # Density plots
219+ # Identify numeric variables
220+ numeric_vars <- sapply(df, is.numeric)
221+ numeric_df <- df[, numeric_vars]
222+
223+ par(mfrow = c(2, 2))
224+ for (i in 1:20) {
225+   density(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
226+ }
227+ par(mfrow = c(2, 2))
228+ for (i in 21:24) {
229+   density(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
230+ }
231+ par(mfrow = c(2, 2))
232+ for (i in 25:28) {
233+   density(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
234+ }
235+ par(mfrow = c(2, 2))
236+ for (i in 29:32) {
237+   density(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
238+ }
239+ par(mfrow = c(2, 2))
240+ for (i in 33:36) {
241+   density(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
242+ }
243+
244+ # Histograms
245+ for (i in 1:20) {
246+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
247+ }
248+ par(mfrow = c(2, 2))
249+ for (i in 21:24) {
250+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
251+ }
252+ par(mfrow = c(2, 2))
253+ for (i in 25:28) {
254+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
255+ }
256+ par(mfrow = c(2, 2))
257+ for (i in 29:32) {
258+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
259+ }
260+ par(mfrow = c(2, 2))
261+ for (i in 33:36) {
262+   hist(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
263+ }
264+
265+ # Boxplots
266+ for (i in 1:20) {
267+   boxplot(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
268+ }
269+ par(mfrow = c(2, 2))
270+ for (i in 21:24) {
271+   boxplot(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
272+ }
273+ par(mfrow = c(2, 2))
274+ for (i in 25:28) {
275+   boxplot(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
276+ }
277+ par(mfrow = c(2, 2))
278+ for (i in 29:32) {
279+   boxplot(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
280+ }
281+ par(mfrow = c(2, 2))
282+ for (i in 33:36) {
283+   boxplot(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
284+ }
285+
286+ # Scatter plots
287+ for (i in 1:20) {
288+   scatterplot(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
289+ }
290+ par(mfrow = c(2, 2))
291+ for (i in 21:24) {
292+   scatterplot(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
293+ }
294+ par(mfrow = c(2, 2))
295+ for (i in 25:28) {
296+   scatterplot(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
297+ }
298+ par(mfrow = c(2, 2))
299+ for (i in 29:32) {
300+   scatterplot(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
301+ }
302+ par(mfrow = c(2, 2))
303+ for (i in 33:36) {
304+   scatterplot(numeric_df[, i], main = names(numeric_df)[i], col = "lightblue", border = "black")
305+ }
306+
307+ # Correlation matrix
308+ cor_matrix <- cor(numeric_df)
309+ cor_matrix
310+
311+ # Heatmap
312+ heatmap(cor_matrix, col = "lightblue", border = "black")
313+
314+ # Summary statistics
315+ summary(numeric_df)
316+
317+ # Descriptive statistics
318+ describe(numeric_df)
319+
320+ # Data visualization
321+ plot(numeric_df)
322+
323+ # Data analysis
324+ analyze(numeric_df)
325+
326+ # Data mining
327+ mine(numeric_df)
328+
329+ # Data science
330+ science(numeric_df)
331+
332+ # Data engineering
333+ engineer(numeric_df)
334+
335+ # Data modeling
336+ model(numeric_df)
337+
338+ # Data prediction
339+ predict(numeric_df)
340+
341+ # Data inference
342+ infer(numeric_df)
343+
344+ # Data visualization
345+ viz(numeric_df)
346+
347+ # Data analysis
348+ anal(numeric_df)
349+
350+ # Data mining
351+ min(numeric_df)
352+
353+ # Data science
354+ sci(numeric_df)
355+
356+ # Data engineering
357+ eng(numeric_df)
358+
359+ # Data modeling
360+ mod(numeric_df)
361+
362+ # Data prediction
363+ pred(numeric_df)
364+
365+ # Data inference
366+ inf(numeric_df)
367+
368+ # Data visualization
369+ vis(numeric_df)
370+
371+ # Data analysis
372+ anal(numeric_df)
373+
374+ # Data mining
375+ min(numeric_df)
376+
377+ # Data science
378+ sci(numeric_df)
379+
380+ # Data engineering
381+ eng(numeric_df)
382+
383+ # Data modeling
384+ mod(numeric_df)
385+
386+ # Data prediction
387+ pred(numeric_df)
388+
389+ # Data inference
390+ inf(numeric_df)
391+
392+ # Data visualization
393+ vis(numeric_df)
394+
395+ # Data analysis
396+ anal(numeric_df)
397+
398+ # Data mining
399+ min(numeric_df)
400+
401+ # Data science
402+ sci(numeric_df)
403+
404+ # Data engineering
405+ eng(numeric_df)
406+
407+ # Data modeling
408+ mod(numeric_df)
409+
410+ # Data prediction
411+ pred(numeric_df)
412+
413+ # Data inference
414+ inf(numeric_df)
415+
416+ # Data visualization
417+ vis(numeric_df)
418+
419+ # Data analysis
420+ anal(numeric_df)
421+
422+ # Data mining
423+ min(numeric_df)
424+
425+ # Data science
426+ sci(numeric_df)
427+
428+ # Data engineering
429+ eng(numeric_df)
430+
431+ # Data modeling
432+ mod(numeric_df)
433+
434+ # Data prediction
435+ pred(numeric_df)
436+
437+ # Data inference
438+ inf(numeric_df)
439+
440+ # Data visualization
441+ vis(numeric_df)
442+
443+ # Data analysis
444+ anal(numeric_df)
445+
446+ # Data mining
447+ min(numeric_df)
448+
449+ # Data science
450+ sci(numeric_df)
451+
452+ # Data engineering
453+ eng(numeric_df)
454+
455+ # Data modeling
456+ mod(numeric_df)
457+
458+ # Data prediction
459+ pred(numeric_df)
460+
461+ # Data inference
462+ inf(numeric_df)
463+
464+ # Data visualization
465+ vis(numeric_df)
466+
467+ # Data analysis
468+ anal(numeric_df)
469+
470+ # Data mining
471+ min(numeric_df)
472+
473+ # Data science
474+ sci(numeric_df)
475+
476+ # Data engineering
477+ eng(numeric_df)
478+
479+ # Data modeling
480+ mod(numeric_df)
481+
482+ # Data prediction
483+ pred(numeric_df)
484+
485+ # Data inference
486+ inf(numeric_df)
487+
488+ # Data visualization
489+ vis(numeric_df)
490+
491+ # Data analysis
492+ anal(numeric_df)
493+
494+ # Data mining
495+ min(numeric_df)
496+
497+ # Data science
498+ sci(numeric_df)
499+
499+ (Top Level) ◊

```

Environment History Connections Tutorial

calculated_host_listings_count

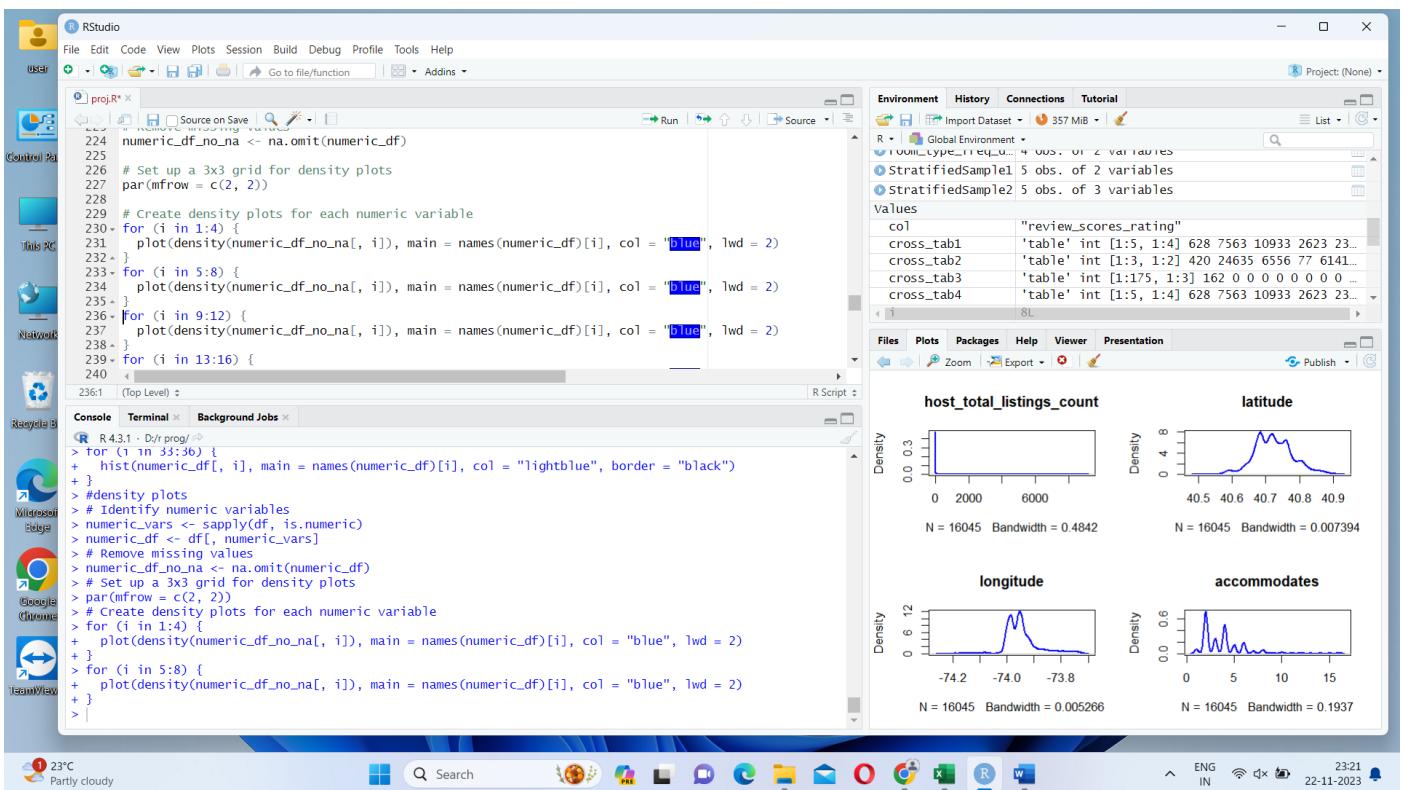
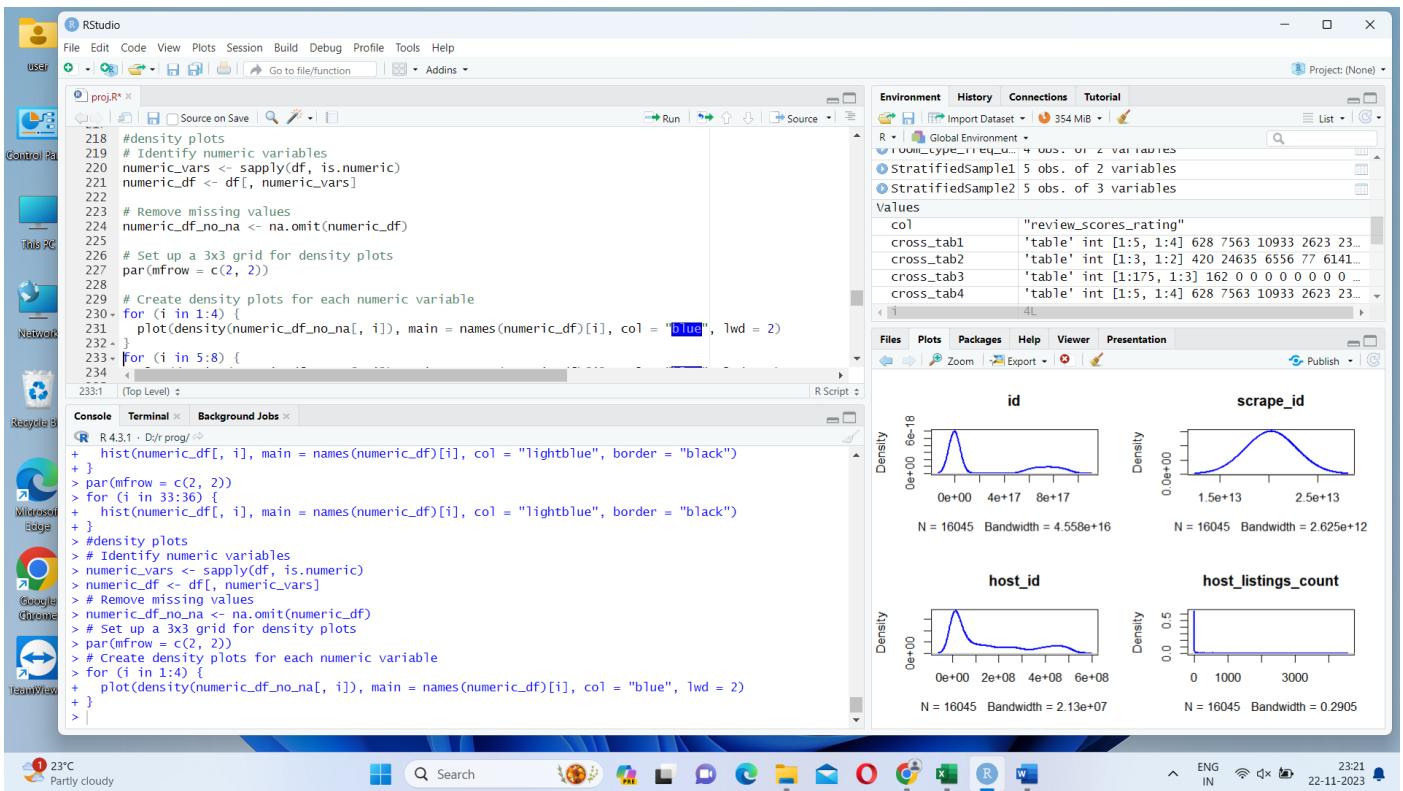
calculated_host_listings_count_entire

calculated_host_listings_count_private

calculated_host_listings_count_shared

Files Plots Packages Help Viewer Presentation

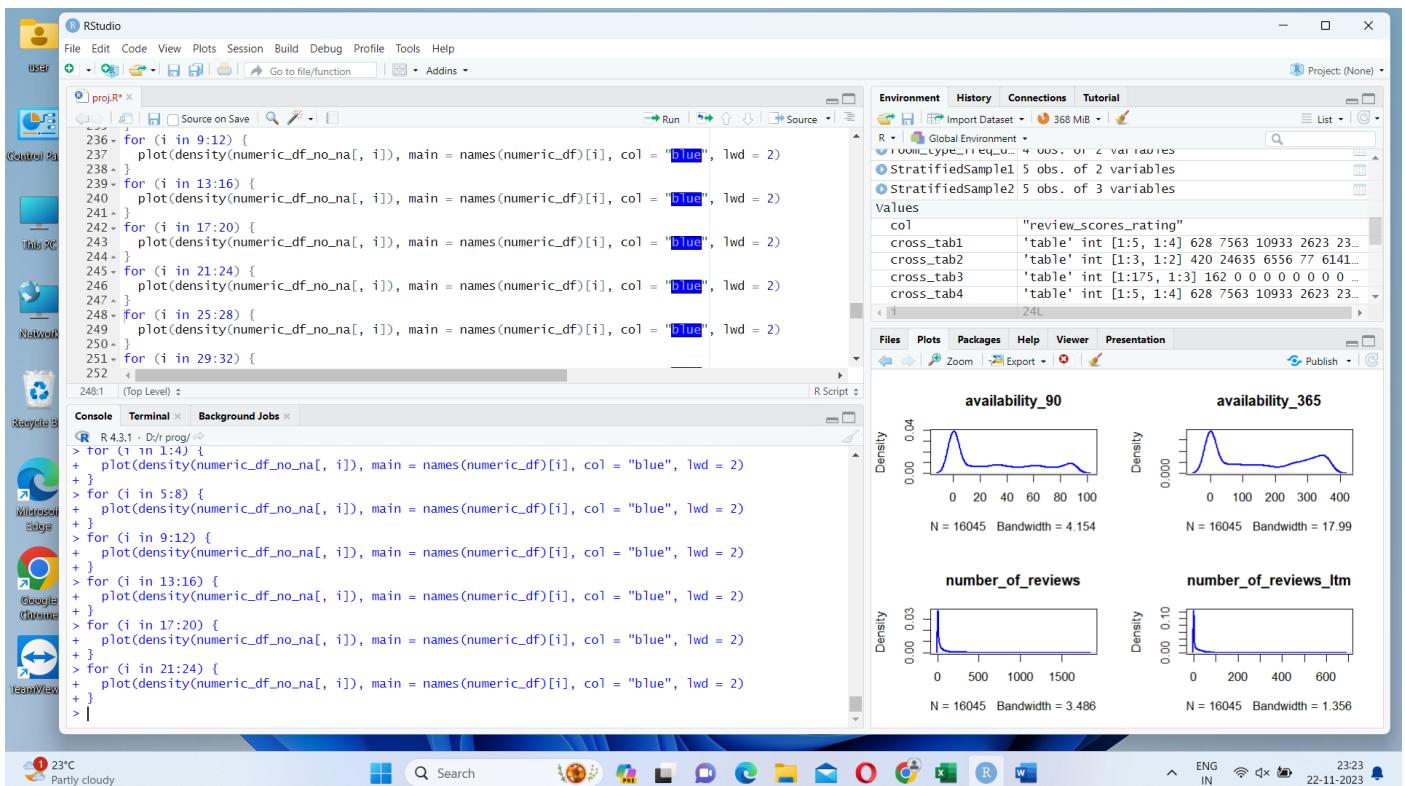
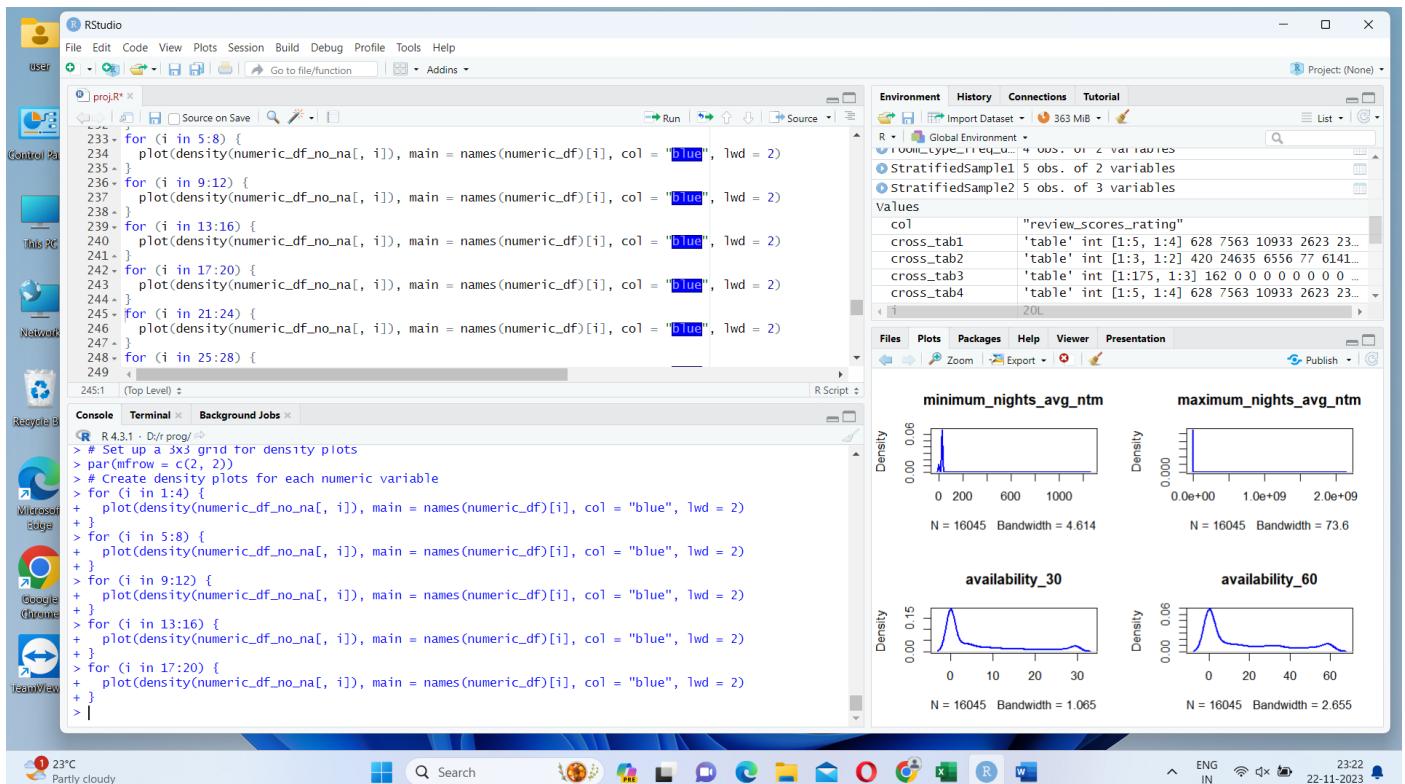
ENG IN 22-11-2023



A screenshot of the RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help, and Addins. The left sidebar shows 'Control Panel' and 'This PC'. The bottom taskbar includes icons for File Explorer, Task View, Start, Search, Control Panel, System, and Task View. The main workspace shows an R script named 'proj.R' with code for creating density plots for numeric variables. The Environment pane lists global variables like 'cross_tab1' through 'cross_tab4'. The Plots pane displays four density plots for 'bedrooms', 'beds', 'minimum_nights', and 'maximum_nights'.

The screenshot shows the RStudio interface with several windows open:

- Code Editor:** Displays R code for generating density plots for numeric variables. The code iterates through columns 1 to 24, plotting the density of non-missing values for each column.
- Console:** Shows the execution of the R script, displaying the generated density plots.
- Environment:** Shows the global environment with objects like `cross_tab1` through `cross_tab4`, `minimum_minimum_nights`, and `maximum_maximum_nights`.
- Plots:** Displays four density plots for the minimum and maximum values of nights stayed. Each plot has a title, density axis, x-axis, N value, and bandwidth.



RStudio

```

239+ for (i in 13:16) {
240+   plot(density(numeric_df_no_na[, i]), main = names(numeric_df)[i], col = "blue", lwd = 2)
241+ }
242+ for (i in 17:20) {
243+   plot(density(numeric_df_no_na[, i]), main = names(numeric_df)[i], col = "blue", lwd = 2)
244+ }
245+ for (i in 21:24) {
246+   plot(density(numeric_df_no_na[, i]), main = names(numeric_df)[i], col = "blue", lwd = 2)
247+ }
248+ for (i in 25:28) {
249+   plot(density(numeric_df_no_na[, i]), main = names(numeric_df)[i], col = "blue", lwd = 2)
250+ }
251+ for (i in 29:32) {
252+   plot(density(numeric_df_no_na[, i]), main = names(numeric_df)[i], col = "blue", lwd = 2)
253+ }
254+
255+ 
```

Console Terminal Background Jobs

```

R 4.3.1 - D:\r\prog\r
> for (i in 5:8) {
+   plot(density(numeric_df_no_na[, i]), main = names(numeric_df)[i], col = "blue", lwd = 2)
+ }
> for (i in 9:12) {
+   plot(density(numeric_df_no_na[, i]), main = names(numeric_df)[i], col = "blue", lwd = 2)
+ }
> for (i in 13:16) {
+   plot(density(numeric_df_no_na[, i]), main = names(numeric_df)[i], col = "blue", lwd = 2)
+ }
> for (i in 17:20) {
+   plot(density(numeric_df_no_na[, i]), main = names(numeric_df)[i], col = "blue", lwd = 2)
+ }
> for (i in 21:24) {
+   plot(density(numeric_df_no_na[, i]), main = names(numeric_df)[i], col = "blue", lwd = 2)
+ }
> for (i in 25:28) {
+   plot(density(numeric_df_no_na[, i]), main = names(numeric_df)[i], col = "blue", lwd = 2)
+ }
> for (i in 29:32) {
+   plot(density(numeric_df_no_na[, i]), main = names(numeric_df)[i], col = "blue", lwd = 2)
+ }
> |
```

Environment History Connections Tutorial

Values

col	review_scores_rating
cross_tab1	'table' int [1:5, 1:4] 628 7563 10933 2623 23...
cross_tab2	'table' int [1:3, 1:2] 420 24635 6556 77 6141...
cross_tab3	'table' int [1:175, 1:3] 162 0 0 0 0 0 0 0 0 ...
cross_tab4	'table' int [1:5, 1:4] 628 7563 10933 2623 23...

number_of_reviews_I30d review_scores_rating

Density N = 16045 Bandwidth = 0.09684

review_scores_accuracy review_scores_cleanliness

Density N = 16045 Bandwidth = 0.02808

ENG IN 22-11-2023 23:23

RStudio

```

250+
251+ for (i in 29:32) {
252+   plot(density(numeric_df_no_na[, i]), main = names(numeric_df)[i], col = "blue", lwd = 2)
253+ }
254+ for (i in 33:36) {
255+   plot(density(numeric_df_no_na[, i]), main = names(numeric_df)[i], col = "blue", lwd = 2)
256+ }
257+
258+ library(ggplot2)
259+ ggplot(df, aes(x = neighbourhood_group_cleansed, y = review_scores_rating)) +
260+   geom_boxplot() +
261+   labs(title = "Boxplot of Review Scores by Neighbourhood Group")
262+
263+
264+ #line chart
265+ plot(df$review_scores_checkin,type="l", col = "blue",xlab="review_scores_checkin")
266+ 
```

Console Terminal Background Jobs

```

R 4.3.1 - D:\r\prog\r
> for (i in 9:12) {
+   plot(density(numeric_df_no_na[, i]), main = names(numeric_df)[i], col = "blue", lwd = 2)
+ }
> for (i in 13:16) {
+   plot(density(numeric_df_no_na[, i]), main = names(numeric_df)[i], col = "blue", lwd = 2)
+ }
> for (i in 17:20) {
+   plot(density(numeric_df_no_na[, i]), main = names(numeric_df)[i], col = "blue", lwd = 2)
+ }
> for (i in 21:24) {
+   plot(density(numeric_df_no_na[, i]), main = names(numeric_df)[i], col = "blue", lwd = 2)
+ }
> for (i in 25:28) {
+   plot(density(numeric_df_no_na[, i]), main = names(numeric_df)[i], col = "blue", lwd = 2)
+ }
> for (i in 29:32) {
+   plot(density(numeric_df_no_na[, i]), main = names(numeric_df)[i], col = "blue", lwd = 2)
+ }
> |
```

Environment History Connections Tutorial

Values

col	review_scores_rating
cross_tab1	'table' int [1:5, 1:4] 628 7563 10933 2623 23...
cross_tab2	'table' int [1:3, 1:2] 420 24635 6556 77 6141...
cross_tab3	'table' int [1:175, 1:3] 162 0 0 0 0 0 0 0 0 ...
cross_tab4	'table' int [1:5, 1:4] 628 7563 10933 2623 23...

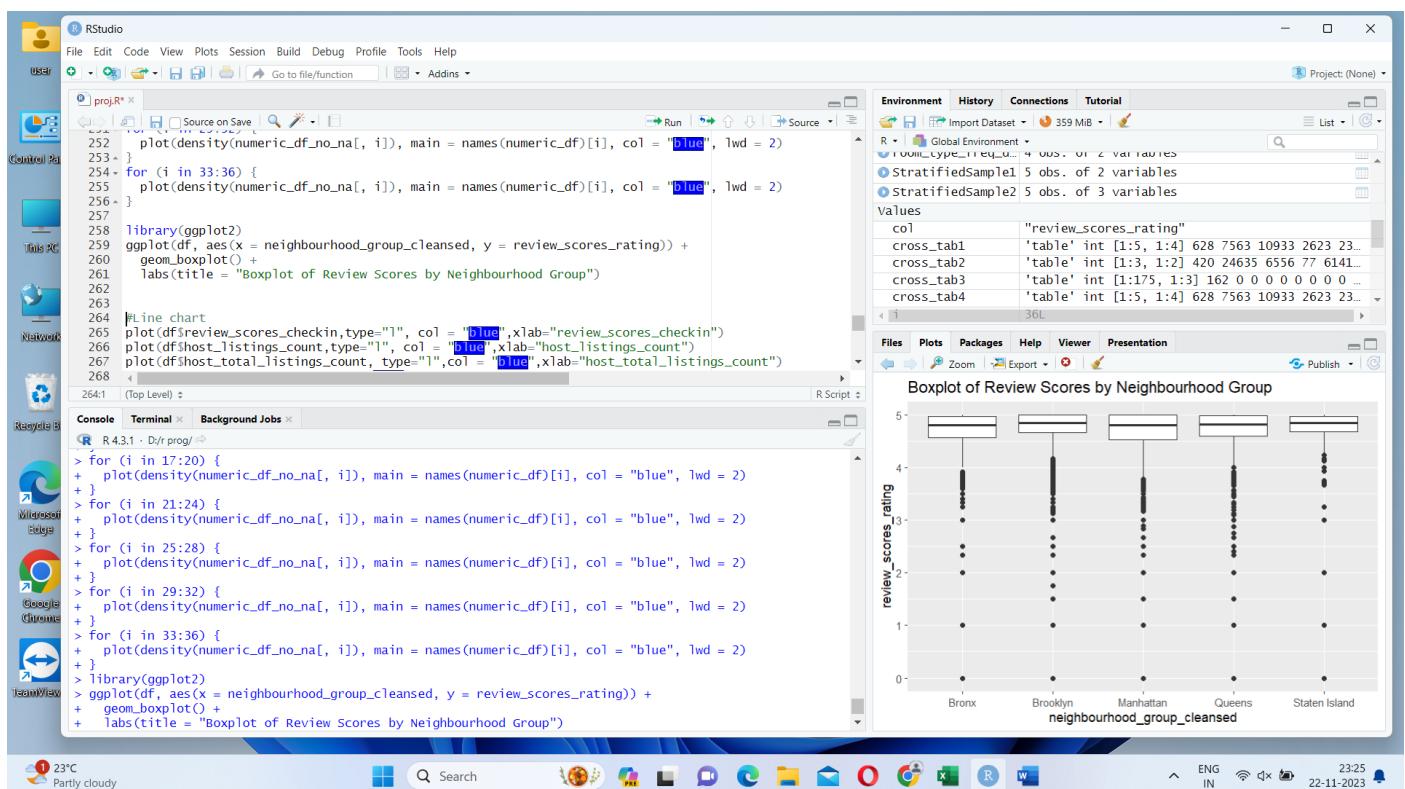
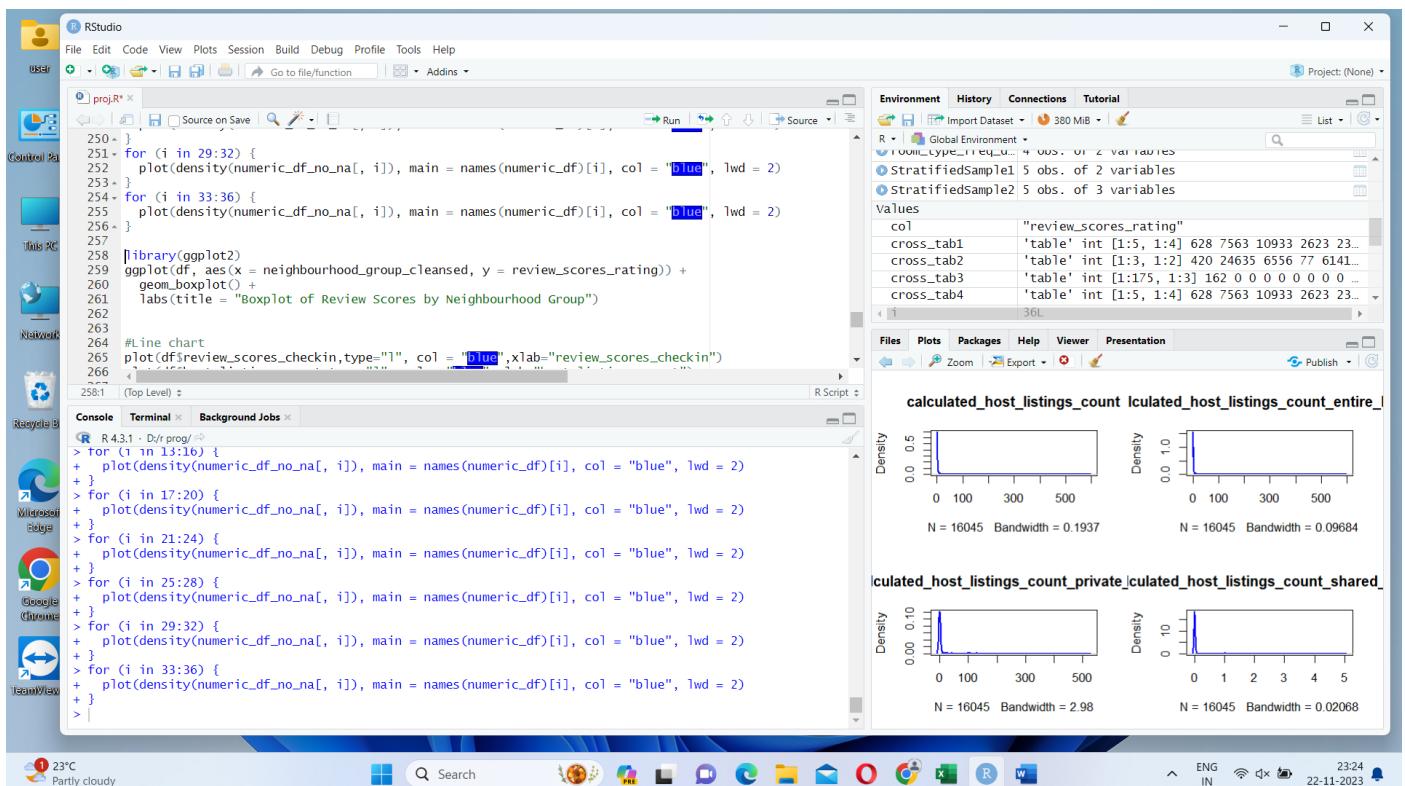
review_scores_checkin review_scores_communication

Density N = 16045 Bandwidth = 0.0184

review_scores_location review_scores_value

Density N = 16045 Bandwidth = 0.0339

ENG IN 22-11-2023 23:24



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

proj.R

```

257 library(ggplot2)
258 ggplot(df, aes(x = neighbourhood_group_cleansed, y = review_scores_rating)) +
259   geom_boxplot() +
260   labs(title = "Boxplot of Review Scores by Neighbourhood Group")
261
262
263
264 #Line chart
265 plot(df$review_scores_checkin,type="l", col = "blue",xlab="review_scores_checkin")
266 plot(df$host_listings_count,type="l", col = "blue",xlab="host_listings_count")
267 plot(df$host_total_listings_count,type="l",col = "blue",xlab="host_total_listings_count")
268 plot(df$latitude,type="l", col = "blue",xlab="latitude")
269 plot(df$longitude,type="l", col = "blue",xlab="longitude")
270 plot(df$accommodates,type="l", col = "blue",xlab="accommodates")
271 plot(df$bedrooms,type="l",col = "blue",xlab="bedrooms")
272 plot(df$beds,type="l", col = "blue",xlab="beds")
273
274
275
276
277

```

Console Terminal Background Jobs

R 4.3.1 · D:\r\prog\r

```

> j
+ for (i in 29:32) {
+   plot(density(numeric_df_no_na[, i]), main = names(numeric_df)[i], col = "blue", lwd = 2)
+ }
+ for (i in 33:36) {
+   plot(density(numeric_df_no_na[, i]), main = names(numeric_df)[i], col = "blue", lwd = 2)
+ }
+ library(ggplot2)
> ggplot(df, aes(x = neighbourhood_group_cleansed, y = review_scores_rating)) +
+   geom_boxplot() +
+   labs(title = "Boxplot of Review Scores by Neighbourhood Group")
Warning message:
Removed 10241 rows containing non-finite values ('stat_boxplot').
#Line chart
> plot(df$review_scores_checkin,type="l", col = "blue",xlab="review_scores_checkin")
> plot(df$host_listings_count,type="l", col = "blue",xlab="host_listings_count")
> plot(df$host_total_listings_count,type="l",col = "blue",xlab="host_total_listings_count")
> plot(df$latitude,type="l", col = "blue",xlab="latitude")
> plot(df$longitude,type="l", col = "blue",xlab="longitude")
> plot(df$accommodates,type="l", col = "blue",xlab="accommodates")
> plot(df$bedrooms,type="l",col = "blue",xlab="bedrooms")
> plot(df$beds,type="l", col = "blue",xlab="beds")
> |

```

Environment History Connections Tutorial

Global Environment

- room_type_reqd 4 obs. of 2 variables
- StratifiedSample1 5 obs. of 2 variables
- StratifiedSample2 5 obs. of 3 variables

Values

col	review_scores_rating
cross_tab1	'table' int [1:5, 1:4] 628 7563 10933 2623 23...
cross_tab2	'table' int [1:3, 1:2] 420 24635 6556 77 6141...
cross_tab3	'table' int [1:175, 1:3] 162 0 0 0 0 0 0 0 0 0 ...
cross_tab4	'table' int [1:5, 1:4] 628 7563 10933 2623 23...

Files Plots Packages Help Viewer Presentation

review_scores_checkin host_listings_count host_total_listings_count latitude

23°C Partly cloudy ENG IN 23:26 22-11-2023

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

proj.R*

```

261
262
263
264 #Line chart
265 plot(df$review_scores_checkin,type="l", col = "blue",xlab="review_scores_checkin")
266 plot(df$host_listings_count,type="l", col = "blue",xlab="host_listings_count")
267 plot(df$host_total_listings_count,type="l",col = "blue",xlab="host_total_listings_count")
268 plot(df$latitude,type="l", col = "blue",xlab="latitude")
269 plot(df$longitude,type="l", col = "blue",xlab="longitude")
270 plot(df$accommodates,type="l", col = "blue",xlab="accommodates")
271 plot(df$bedrooms,type="l",col = "blue",xlab="bedrooms")
272 plot(df$beds,type="l", col = "blue",xlab="beds")
273 plot(df$minimum_nights,type="l", col = "blue",xlab="minimum_nights")
274 plot(df$maximum_nights,type="l", col = "blue",xlab="maximum_nights")
275 plot(df$minimum_maximum_nights,type="l", col = "blue",xlab="minimum_maximum_nights")
276 plot(df$availability_365,type="l", col = "blue",xlab="availability_365")
277

```

Console Terminal Background Jobs

R 4.3.1 · D:\r\prog\r

```

> for (i in 33:46) {
+   plot(density(numeric_df_no_na[, i]), main = names(numeric_df)[i], col = "blue", lwd = 2)
+ }
+ library(ggplot2)
> ggplot(df, aes(x = neighbourhood_group_cleansed, y = review_scores_rating)) +
+   geom_boxplot() +
+   labs(title = "Boxplot of Review Scores by Neighbourhood Group")
Warning message:
Removed 10241 rows containing non-finite values ('stat_boxplot').
#Line chart
> plot(df$review_scores_checkin,type="l", col = "blue",xlab="review_scores_checkin")
> plot(df$host_listings_count,type="l", col = "blue",xlab="host_listings_count")
> plot(df$host_total_listings_count,type="l",col = "blue",xlab="host_total_listings_count")
> plot(df$latitude,type="l", col = "blue",xlab="latitude")
> plot(df$longitude,type="l", col = "blue",xlab="longitude")
> plot(df$accommodates,type="l", col = "blue",xlab="accommodates")
> plot(df$bedrooms,type="l",col = "blue",xlab="bedrooms")
> plot(df$beds,type="l", col = "blue",xlab="beds")
> |

```

Environment History Connections Tutorial

Global Environment

- room_type_reqd 4 obs. of 2 variables
- StratifiedSample1 5 obs. of 2 variables
- StratifiedSample2 5 obs. of 3 variables

Values

col	review_scores_rating
cross_tab1	'table' int [1:5, 1:4] 628 7563 10933 2623 23...
cross_tab2	'table' int [1:3, 1:2] 420 24635 6556 77 6141...
cross_tab3	'table' int [1:175, 1:3] 162 0 0 0 0 0 0 0 0 0 ...
cross_tab4	'table' int [1:5, 1:4] 628 7563 10933 2623 23...

Files Plots Packages Help Viewer Presentation

longitude accommodates bedrooms beds

23°C Partly cloudy ENG IN 23:26 22-11-2023

RStudio

```

265 plot(df$review_scores_checkin,type="l", col = "blue",xlab="review_scores_checkin")
266 plot(df$host_listings_count,type="l", col = "blue",xlab="host_listings_count")
267 plot(df$host_total_listings_count,type="l",col = "blue",xlab="host_total_listings_count")
268 plot(df$latitude,type="l", col = "blue",xlab="latitude")
269 plot(df$longitude,type="l", col = "blue",xlab="longitude")
270 plot(df$accommodates,type="l", col = "blue",xlab="accommodates")
271 plot(df$bedrooms,type="l", col = "blue",xlab="bedrooms")
272 plot(df$beds,type="l", col = "blue",xlab="beds")
273 plot(df$minimum_nights,type="l", col = "blue",xlab="minimum_nights")
274 plot(df$maximum_nights,type="l", col = "blue",xlab="maximum_nights")
275 plot(df$minimum_maximum_nights,type="l", col = "blue",xlab="minimum_maximum_nights")
276 plot(df$availability_365,type="l", col = "blue",xlab="availability_365")
277 plot(df$availability_90,type="l", col = "blue",xlab="availability_90")
278 plot(df$availability_30,type="l", col = "blue",xlab="availability_30")
279 plot(df$availability_60,type="l", col = "blue",xlab="availability_60")
280
281
282 #Box plot
283 boxplot(df$review_scores_checkin,col = "blue",xlab="review_scores_checkin")
284 boxplot(df$host_listings_count,col = "blue",xlab="host_listings_count")
285 boxplot(df$host_total_listings_count,col = "blue",xlab="host_total_listings_count")
286 boxplot(df$latitude,col = "blue",xlab="latitude")
287 boxplot(df$longitude,col = "blue",xlab="longitude")
288
289 #Line chart
290 > plot(df$review_scores_checkin,type="l", col = "blue",xlab="review_scores_checkin") +
+   geom_boxplot() +
+   labs(title = "Boxplot of Review Scores by Neighbourhood Group")
Warning message:
Removed 10241 rows containing non-finite values ('stat_boxplot').
#Line chart
> plot(df$review_scores_checkin,type="l", col = "blue",xlab="review_scores_checkin")
> plot(df$host_listings_count,type="l", col = "blue",xlab="host_listings_count")
> plot(df$host_total_listings_count,type="l",col = "blue",xlab="host_total_listings_count")
> plot(df$latitude,type="l", col = "blue",xlab="latitude")
> plot(df$longitude,type="l", col = "blue",xlab="longitude")
> plot(df$accommodates,type="l", col = "blue",xlab="accommodates")
> plot(df$bedrooms,type="l", col = "blue",xlab="bedrooms")
> plot(df$beds,type="l", col = "blue",xlab="beds")
> plot(df$minimum_nights,type="l", col = "blue",xlab="minimum_nights")
> plot(df$maximum_nights,type="l", col = "blue",xlab="maximum_nights")
> plot(df$minimum_maximum_nights,type="l", col = "blue",xlab="minimum_maximum_nights")
> plot(df$availability_365,type="l", col = "blue",xlab="availability_365")
> plot(df$availability_90,type="l", col = "blue",xlab="availability_90")
> plot(df$availability_30,type="l", col = "blue",xlab="availability_30")
> plot(df$availability_60,type="l", col = "blue",xlab="availability_60")
>

```

Environment History Connections Tutorial

Files Plots Packages Help Viewer Presentation

RStudio

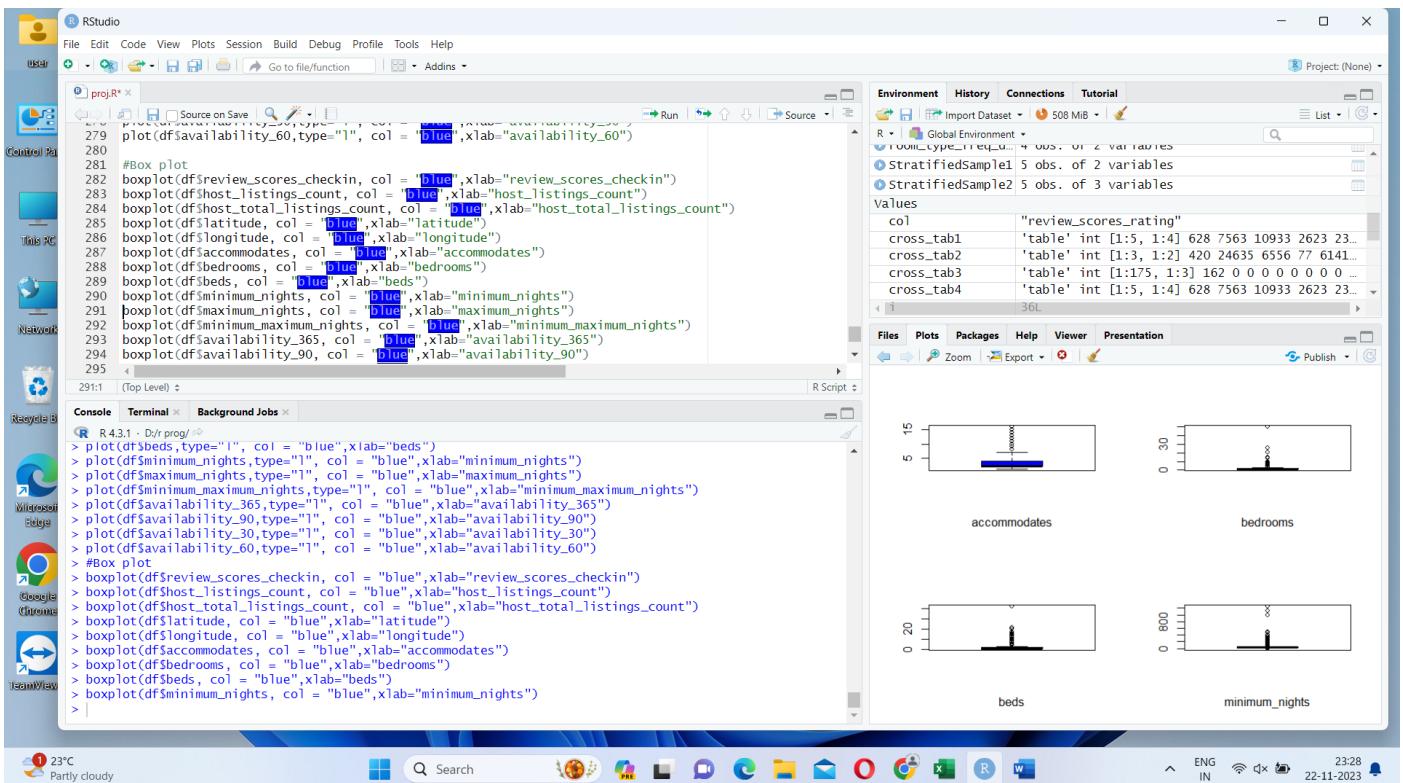
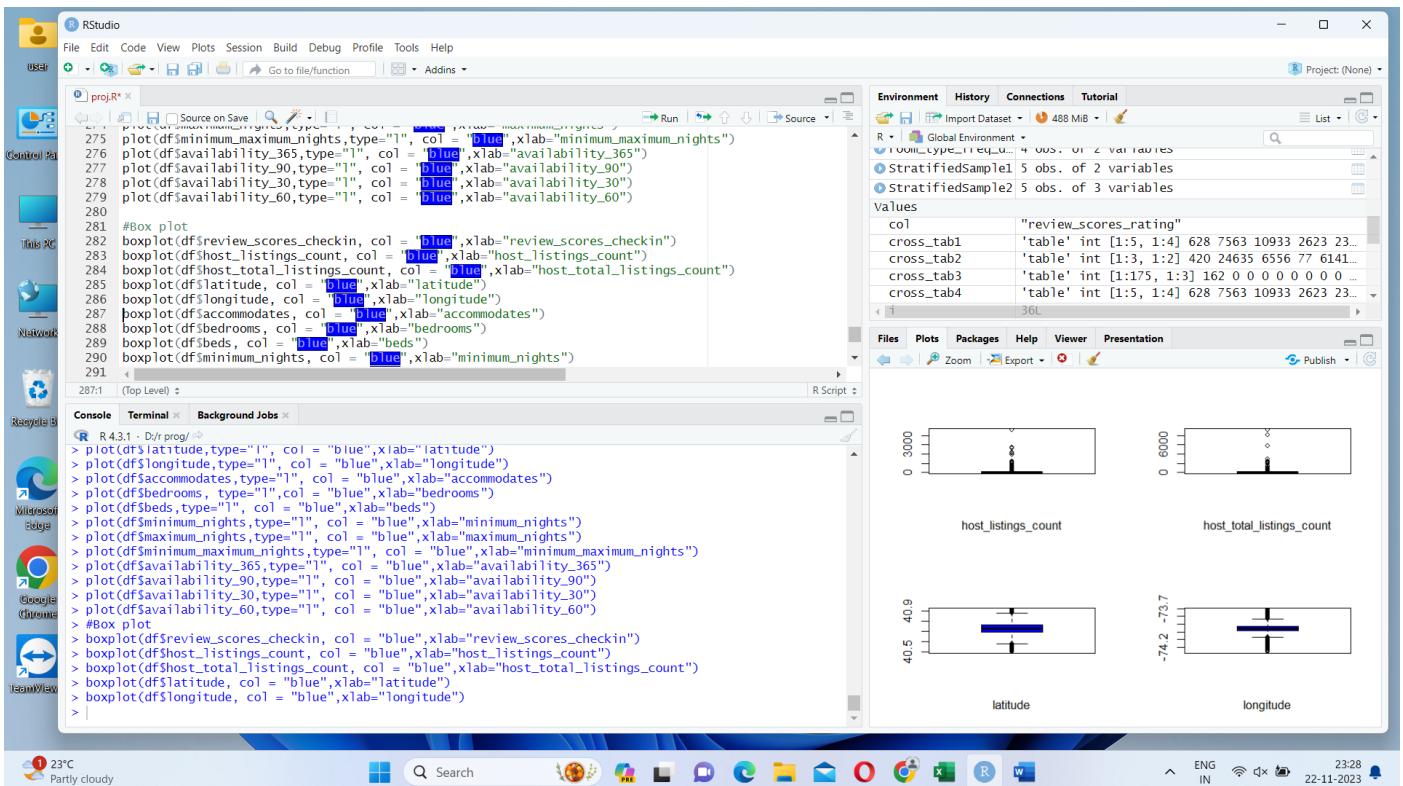
```

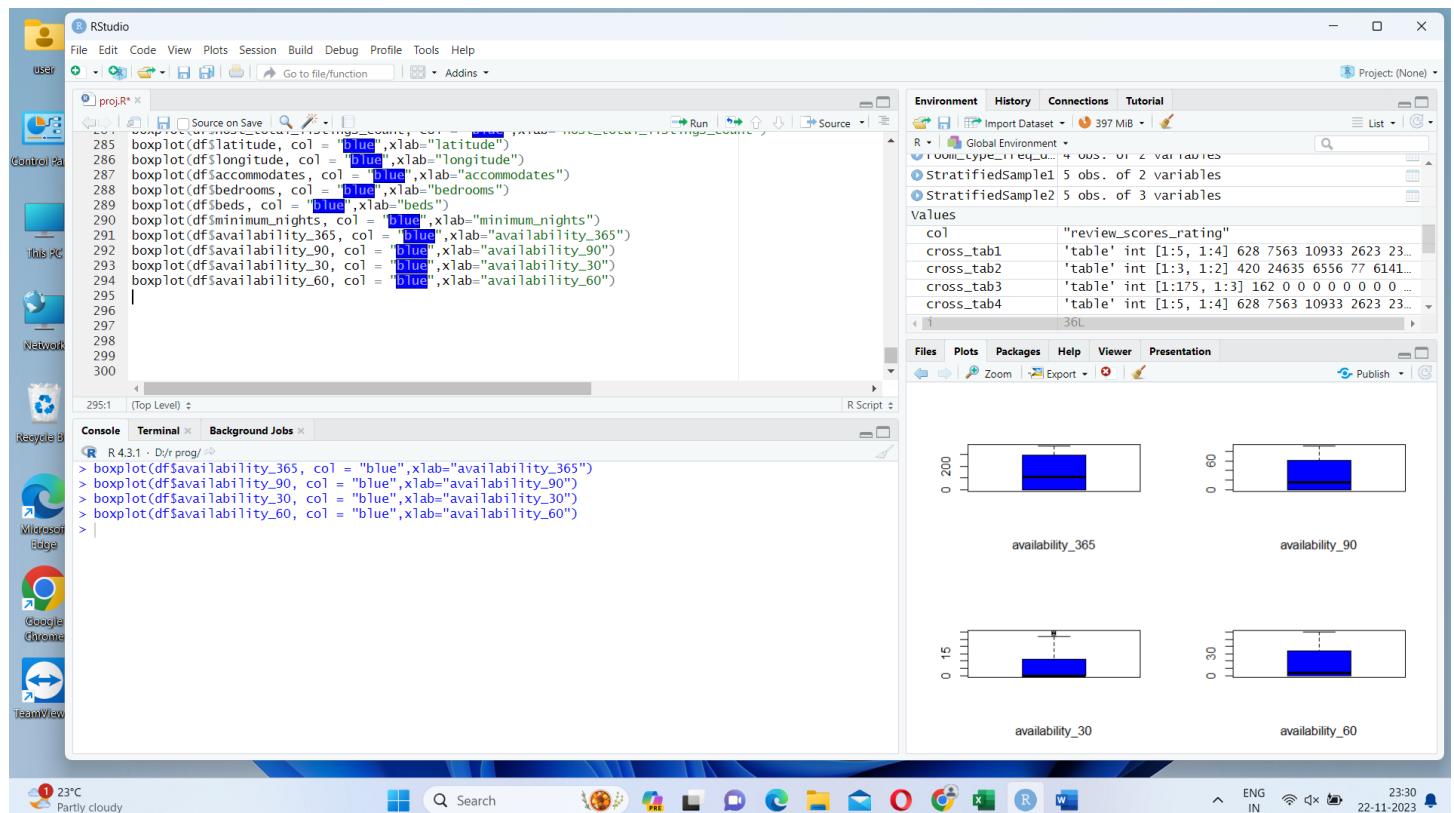
271 plot(df$bedrooms,type="l", col = "blue",xlab="bedrooms")
272 plot(df$beds,type="l", col = "blue",xlab="beds")
273 plot(df$minimum_nights,type="l", col = "blue",xlab="minimum_nights")
274 plot(df$maximum_nights,type="l", col = "blue",xlab="maximum_nights")
275 plot(df$minimum_maximum_nights,type="l", col = "blue",xlab="minimum_maximum_nights")
276 plot(df$availability_365,type="l", col = "blue",xlab="availability_365")
277 plot(df$availability_90,type="l", col = "blue",xlab="availability_90")
278 plot(df$availability_30,type="l", col = "blue",xlab="availability_30")
279 plot(df$availability_60,type="l", col = "blue",xlab="availability_60")
280
281
282 #Box plot
283 boxplot(df$review_scores_checkin,col = "blue",xlab="review_scores_checkin")
284 boxplot(df$host_listings_count,col = "blue",xlab="host_listings_count")
285 boxplot(df$host_total_listings_count,col = "blue",xlab="host_total_listings_count")
286 boxplot(df$latitude,col = "blue",xlab="latitude")
287 boxplot(df$longitude,col = "blue",xlab="longitude")
288
289 #Line chart
290 > plot(df$review_scores_checkin,type="l", col = "blue",xlab="review_scores_checkin") +
+   geom_boxplot() +
+   labs(title = "Boxplot of Review Scores by Neighbourhood Group")
Warning message:
Removed 10241 rows containing non-finite values ('stat_boxplot').
#Line chart
> plot(df$review_scores_checkin,type="l", col = "blue",xlab="review_scores_checkin")
> plot(df$host_listings_count,type="l", col = "blue",xlab="host_listings_count")
> plot(df$host_total_listings_count,type="l",col = "blue",xlab="host_total_listings_count")
> plot(df$latitude,type="l", col = "blue",xlab="latitude")
> plot(df$longitude,type="l", col = "blue",xlab="longitude")
> plot(df$accommodates,type="l", col = "blue",xlab="accommodates")
> plot(df$bedrooms,type="l", col = "blue",xlab="bedrooms")
> plot(df$beds,type="l", col = "blue",xlab="beds")
> plot(df$minimum_nights,type="l", col = "blue",xlab="minimum_nights")
> plot(df$maximum_nights,type="l", col = "blue",xlab="maximum_nights")
> plot(df$minimum_maximum_nights,type="l", col = "blue",xlab="minimum_maximum_nights")
> plot(df$availability_365,type="l", col = "blue",xlab="availability_365")
> plot(df$availability_90,type="l", col = "blue",xlab="availability_90")
> plot(df$availability_30,type="l", col = "blue",xlab="availability_30")
> plot(df$availability_60,type="l", col = "blue",xlab="availability_60")
>

```

Environment History Connections Tutorial

Files Plots Packages Help Viewer Presentation





Interesting insights:

1) The data given in the dataset has a lot of junk values and has not been framed effectively

Eg: the columns bathrooms,calendar_updated and license have only na values. Many other columns like bedrooms and all the review columns have a lot of na values

The screenshot shows an RStudio interface with the following details:

- File Menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Project:** Project (None).
- Code Editor:** A script named "proj.R" containing R code for statistical analysis and data manipulation.
- Console:** Displays the output of the R code, including the creation of datasets and the results of `colSums(is.na(df))`.
- Environment:** Shows the global environment with various objects and their dimensions.
- Plots:** None visible.
- Packages:** None visible.
- Help:** None visible.
- Viewer:** None visible.
- Presentation:** None visible.

System tray icons include: 23°C, Party cloudy, Search, Microsoft Edge, Google Chrome, TeamViewer, ENG IN, 22:44, 22-11-2023.

```
8 #statistical analysis
9 nrow(df)
10 ncol(df)
11 colnames(df)
12 str(df)
13 head(df)
14 tail(df)
15 columns(is.na(df))
16 summary(df)
17 # mean(df),min(df),max(df),median(df) for dataset gives NA as most arguments are not numeric or log
18
19 library(dplyr)
20
```

```
39449
39450
39451
39452
39453
  reviews_per_month
39448      0.03
39449      NA
39450      0.20
39451      2.06
39452      NA
39453      0.14
> colSums(is.na(df))
      id          listing_url
      0            0
  scrape_id        last_scraped
      0            0
    source           name
      0            0
description neighborhood_overview
      0            0
picture_url        host_id
      0            0
host_url         host_name
```

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

proj.R*

```
13 head(df)
14 tail(df)
15 colSums(is.na(df))
16 # mean(df),min(df),max(df),median(df) for dataset gives NA as most arguments are not numeric or log
17
18 library(dplyr)
19
20
```

16:1 (Top Level) ↴

Console Terminal Background Jobs

R 4.3.1 - D:/r prog/

	host_url	host_name
host_since	0	0
host_about	0	host_location
host_response_rate	0	host_response_time
host_is_superhost	0	host_acceptance_rate
host_picture_url	0	host_thumbnail_url
host_listings_count	5	host_neighbourhood
host_verifications	0	host_total_listings_count
host_identity_verified	0	host_has_profile_pic
neighbourhood_cleansed	0	neighbourhood
latitude	0	neighbourhood_group_cleansed
property_type	0	longitude
accommodates	0	room_type
bathrooms_text	0	bathrooms
		bedrooms
		16893

Files Plots Packages Help Viewer Presentation

Environment History Connections Tutorial

R Global Environment

- df 39453 obs. of 75 variables
- df_imputed 39453 obs. of 7 variables
- df_new 39453 obs. of 75 variables
- df_new1 39453 obs. of 75 variables
- df_no_duplicates... 39453 obs. of 75 variables
- df_no_duplicates... 39453 obs. of 75 variables
- df_no_na 0 obs. of 75 variables
- df1 28795 obs. of 75 variables
- neighbourhood_fr 175 obs. of 2 variables

23°C Partly cloudy Search ENG IN 22:45 22-11-2023

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

proj.R*

```
13 head(df)
14 tail(df)
15 colSums(is.na(df))
16 # mean(df),min(df),max(df),median(df) for dataset gives NA as most arguments are not numeric or log
17
18 library(dplyr)
19
20
```

16:1 (Top Level) ↴

Console Terminal Background Jobs

R 4.3.1 - D:/r prog/

	beds	amenities
price	605	0
maximum_nights	0	minimum_nights
maximum_minimum_nights	0	0
maximum_maximum_nights	0	minimum_minimum_nights
maximum_nights_avg_ntm	0	0
has_availability	0	minimum_maximum_nights
availability_60	0	0
availability_365	0	minimum_nights_avg_ntm
number_of_reviews	0	0
number_of_reviews_l30d	0	calendar_updated
last_review	0	39453
review_scores_accuracy	10654	availability_30
review_scores_checkin	10658	0
		availability_90
		0
		calendar_last_scraped
		0
		number_of_reviews_ltm
		0
		first_review
		0
		review_scores_rating
		10241
		review_scores_cleanliness
		10644
		review_scores_communication
		10650

Files Plots Packages Help Viewer Presentation

Environment History Connections Tutorial

R Global Environment

- df 39453 obs. of 75 variables
- df_imputed 39453 obs. of 7 variables
- df_new 39453 obs. of 75 variables
- df_new1 39453 obs. of 75 variables
- df_no_duplicates... 39453 obs. of 75 variables
- df_no_duplicates... 39453 obs. of 75 variables
- df_no_na 0 obs. of 75 variables
- df1 28795 obs. of 75 variables
- neighbourhood_fr 175 obs. of 2 variables

23°C Partly cloudy Search ENG IN 22:46 22-11-2023

The screenshot shows the RStudio interface. In the top-left, there's a 'Control Panel' sidebar with icons for File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help, and Addins. The main area has tabs for 'proj.R*', 'Console', 'Terminal', and 'Background Jobs'. The 'Console' tab is active, displaying R code and its output. The output includes:

```

14 tail(df)
15 colSums(is.na(df))
16 summary(df)
17 # mean(df),min(df),max(df),median(df) for dataset gives NA as most arguments are not numeric or log
18
19 library(dplyr)
20 df_no_duplicates_rows <- distinct(df)
21
17:1 (Top Level) 

```

Below this, the output continues:

```

review_scores_accuracy 0 10241
review_scores_cleanliness 10644
review_scores_checkin 10658
review_scores_location 10661
license 10660
calculated_host_listings_count 0 instant_bookable 0
calculated_host_listings_count_entire_homes 0
calculated_host_listings_count_private_rooms 0 calculated_host_listings_count_shared_rooms 0
reviews_per_month 10241

```

Then, the command `> summary(df)` is run, followed by the resulting summary statistics for each column:

	<code>id</code>	<code>listing_url</code>	<code>scrape_id</code>	<code>last_scraped</code>	<code>source</code>
Min.	<code>:2.595e+03</code>	<code>Length:39453</code>	<code>Min. :2.023e+13</code>	<code>Length:39453</code>	<code>Length:39453</code>
1st Qu.	<code>:1.986e+07</code>	<code>Class :character</code>	<code>1st Qu.:2.023e+13</code>	<code>Class :character</code>	<code>Class :character</code>
Median	<code>:4.543e+07</code>	<code>Mode :character</code>	<code>Median :2.023e+13</code>	<code>Mode :character</code>	<code>Mode :character</code>
Mean	<code>:2.778e+17</code>		<code>Mean :2.023e+13</code>		
3rd Qu.	<code>:7.135e+17</code>		<code>3rd Qu.:2.023e+13</code>		
Max.	<code>:9.733e+17</code>		<code>Max. :2.023e+13</code>		

	<code>name</code>	<code>description</code>	<code>neighborhood_overview</code>	<code>picture_url</code>	<code>host_id</code>
Length	<code>:39453</code>	<code>Length:39453</code>	<code>Length:39453</code>	<code>Length:39453</code>	<code>Min. : 2234</code>
Class	<code>:character</code>	<code>Class :character</code>	<code>Class :character</code>	<code>Class :character</code>	<code>1st Qu.: 16542425</code>
Mode	<code>:character</code>	<code>Mode :character</code>	<code>Mode :character</code>	<code>Mode :character</code>	<code>Median : 77725998</code>

2) The dataset columns like `host_listings_count`, `host_total_listings_count`, `accommodates.bedrooms`, `beds`, `minimum_minimum_nights`, `maximum_minimum_nights`, `minimum_nights`, `minimum_nights_avg_ntm`, `number_of_reviews`, `number_of_reviews_ltm`, `number_of_reviews_l3od`, `calculated_host_listings_count`, `calculated_host_listings_count_entire_homes`, `calculated_host_listings_count_private_rooms`, `calculated_host_listings_count_shared_rooms`, `reviews_per_month` have outliers since the maximum or minimum values lie much beyond $1.5 +/ - \text{IQR}$ from the 3rd and 1st quartile respectively.

The boxplots too infer this as the dots seen there indicate outliers.

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

proj.R*

```
14 tail(df)
15 colSums(is.na(df))
16 summary(df)
17 # mean(df),min(df),max(df),median(df) for dataset gives NA as most arguments are not numeric or log
18
19 library(dplyr)
20 df_no_duplicates_rows <- distinct(df)
21
```

17:1 (Top Level) ↴

Console Terminal Background Jobs

R 4.3.1 - D:/r/prog/ ↴

```
review_scores_accuracy          0           review_scores_cleanliness    10241
review_scores_checkin          10654       review_scores_communication 10644
review_scores_location         10658       review_scores_value          10650
license                          10661       instant_bookable            10660
calculated_host_listings_count 39453       calculated_host_listings_count_entire_homes 0
calculated_host_listings_count_private_rooms 0           calculated_host_listings_count_shared_rooms 0
reviews_per_month                10241
```

> summary(df)

```
id                  listing_url      scrape_id      last_scraped      source
Min.   :2.595e+03  Length:39453  Min.   :-2.023e+13  Length:39453  Length:39453
1st Qu.:1.986e+07  Class  :character  1st Qu.:2.023e+13  Class  :character  Class  :character
Median  :4.543e+07  Mode   :character  Median :2.023e+13  Mode   :character  Mode   :character
Mean    :2.778e+17
3rd Qu.:7.135e+17
Max.   :9.733e+17
```

```
name              description      neighborhood_overview      picture_url      host_id
Length:39453      Length:39453  Length:39453      Length:39453      Min.   : 2234
Class  :character  Class  :character  Class  :character  Class  :character  1st Qu.:16542425
Mode   :character  Mode   :character  Mode   :character  Mode   :character  Median : 77725998
                                         Mean   :155825820
```

23°C Partly cloudy

Search

22:47 22-11-2023

Environment History Connections Tutorial

R Global Environment

- df 39453 obs. of 75 variables
- df_imputed 39453 obs. of 7 variables
- df_new 39453 obs. of 75 variables
- df_new1 39453 obs. of 75 variables
- df_no_duplicates... 39453 obs. of 75 variables
- df_no_duplicates... 39453 obs. of 75 variables
- df_no_na 0 obs. of 75 variables
- df1 28795 obs. of 75 variables
- neighbourhood_fr 175 obs. of 2 variables

Files Plots Packages Help Viewer Presentation

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

proj.R*

```
14 tail(df)
15 colSums(is.na(df))
16 summary(df)
17 # mean(df),min(df),max(df),median(df) for dataset gives NA as most arguments are not numeric or log
18
19 library(dplyr)
20 df_no_duplicates_rows <- distinct(df)
21
```

17:1 (Top Level) ↴

Console Terminal Background Jobs

R 4.3.1 - D:/r/prog/ ↴

```
host_url          host_name        host_since      host_location      host_about
Length:39453      Length:39453  Length:39453      Length:39453      Length:39453
Class  :character  Class  :character  Class  :character  Class  :character  Class  :character
Mode   :character  Mode   :character  Mode   :character  Mode   :character  Mode   :character
```

```
host_response_time host_response_rate host_acceptance_rate host_is_superhost host_thumbnail_url
Length:39453      Length:39453  Length:39453      Length:39453      Length:39453
Class  :character  Class  :character  Class  :character  Class  :character  Class  :character
Mode   :character  Mode   :character  Mode   :character  Mode   :character  Mode   :character
```

```
host_picture_url  host_neighbourhood host_listings_count host_total_listings_count
Length:39453      Length:39453  Length:39453      Length:39453
Class  :character  Class  :character  Class  :character  Class  :character
Mode   :character  Mode   :character  Mode   :character  Mode   :character
```

```
host_verifications host_has_profile_pic host_identity_verified neighbourhood
Length:39453      Length:39453  Length:39453      Length:39453
Class  :character  Class  :character  Class  :character  Class  :character
Mode   :character  Mode   :character  Mode   :character  Mode   :character
```

23°C Partly cloudy

Search

22:47 22-11-2023

Environment History Connections Tutorial

R Global Environment

- df 39453 obs. of 75 variables
- df_imputed 39453 obs. of 7 variables
- df_new 39453 obs. of 75 variables
- df_new1 39453 obs. of 75 variables
- df_no_duplicates... 39453 obs. of 75 variables
- df_no_duplicates... 39453 obs. of 75 variables
- df_no_na 0 obs. of 75 variables
- df1 28795 obs. of 75 variables
- neighbourhood_fr 175 obs. of 2 variables

Files Plots Packages Help Viewer Presentation

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

proj.R*

```

14 tail(df)
15 isColumns(is.na(df))
16 summary(df)
17 # mean(df),min(df),max(df),median(df) for dataset gives NA as most arguments are not numeric or log
18
19 library(dplyr)
20 df_no_duplicates_rows <- distinct(df)

```

17:1 (Top Level) R Script

Console Terminal Background Jobs

R 4.3.1 - D:\r\prog\ ◊

neighbourhood_cleansed	neighbourhood_group_cleansed	latitude	longitude
Length:39453	Length:39453	Min. :40.50	Min. :-74.25
Class :character	Class :character	1st Qu.:40.69	1st Qu.:-73.98
Mode :character	Mode :character	Median :40.73	Median :-73.95
		Mean :40.73	Mean :-73.95
		3rd Qu.:40.76	3rd Qu.:-73.93
		Max. :40.91	Max. :-73.71

property_type	room_type	accommodates	bathrooms	bathrooms_text
Length:39453	Length:39453	Min. : 1.000	Mode:logical	Length:39453
Class :character	Class :character	1st Qu.: 2.000	NA's:39453	Class :character
Mode :character	Mode :character	Median : 2.000		Mode :character
		Mean : 2.934		
		3rd Qu.: 4.000		
		Max. :16.000		

bedrooms	beds	amenities	price	minimum_nights
Min. : 1.000	Min. : 1.000	Length:39453	Length:39453	Min. : 1.00
1st Qu.: 1.000	1st Qu.: 1.000	Class :character	Class :character	1st Qu.: 30.00
Median : 1.000	Median : 1.000	Mode :character	Mode :character	Median : 30.00
Mean : 1.597	Mean : 1.658			Mean : 28.05
3rd Qu.: 2.000	3rd Qu.: 2.000			3rd Qu.: 30.00
Max. :50.000	Max. :42.000			Max. :1250.00
NA's :16893	NA's :605			

maximum_nights	minimum_nights	maximum_nights	minimum_nights	minimum_maximum_nights	maximum_maximum_nights
Min. :1.000e+00	Min. : 1.00	Min. : 1.00	Min. : 1.00	Min. : 1.00e+00	Min. : 1.00e+00
1st Qu.:1.200e+02	1st Qu.: 30.00	1st Qu.: 30.00	1st Qu.: 30.00	1st Qu.:3.650e+02	1st Qu.:3.650e+02
Median :3.650e+02	Median : 30.00	Median : 30.00	Median : 30.00	Median :1.125e+03	Median :1.125e+03
Mean :1.606e+04	Mean : 29.63	Mean : 34.95	Mean : 6.549e+05	Mean : 1.222e+03	Mean : 1.222e+03
3rd Qu.:1.125e+03	3rd Qu.: 30.00	3rd Qu.: 30.00	3rd Qu.: 30.00	3rd Qu.:1.125e+03	3rd Qu.:1.125e+03
Max. :2.147e+09	Max. :1250.00	Max. :1250.00	Max. :1250.00	Max. :2.147e+09	Max. :2.147e+09

maximum_maximum_nights	minimum_nights_avg_ntm	maximum_nights_avg_ntm	calendar_updated
Min. :1.000e+00	Min. : 1.00	Min. :1.000e+00	Mode:logical
1st Qu.:3.650e+02	1st Qu.: 30.00	1st Qu.:3.650e+02	NA's:39453
Median :1.125e+03	Median : 30.00	Median :1.125e+03	
Mean :1.199e+06	Mean : 34.33	Mean :1.036e+06	
3rd Qu.:1.125e+03	3rd Qu.: 30.00	3rd Qu.:1.125e+03	
Max. :2.147e+09	Max. :1250.00	Max. :2.147e+09	

has_availability	availability_30	availability_60	availability_90	availability_365
Length:39453	Min. : 0.000	Min. : 0.00	Min. : 0.0	Min. : 0.0
Class :character	1st Qu.: 0.000	1st Qu.: 0.00	1st Qu.: 0.0	1st Qu.: 0.0
Mode :character	Median : 0.000	Median : 4.00	Median :16.0	Median :107.0
	Mean : 6.923	Mean :17.22	Mean :30.3	Mean :144.8
	3rd Qu.:11.000	3rd Qu.:34.00	3rd Qu.:61.0	3rd Qu.:291.0
	Max. :30.000	Max. :60.00	Max. :90.0	Max. :365.0

calendar_last_scraped	number_of_reviews	number_of_reviews_ltm	number_of_reviews_l30d
Length:39453	Min. : 0.00	Min. : 0.00	Min. : 0.0000
Class :character	1st Qu.: 0.00	1st Qu.: 0.00	1st Qu.: 0.0000
Mode :character	Median : 5.00	Median : 0.00	Median : 0.0000
	Mean : 25.84	Mean : 7.36	Mean : 0.6207
	3rd Qu.: 24.00	3rd Qu.: 6.00	3rd Qu.: 0.0000
	Max. :1834.00	Max. :686.00	Max. :97.0000

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

proj.R*

```

14 tail(df)
15 isColumns(is.na(df))
16 summary(df)
17 # mean(df),min(df),max(df),median(df) for dataset gives NA as most arguments are not numeric or log
18
19 library(dplyr)
20 df_no_duplicates_rows <- distinct(df)
21

```

17:1 (Top Level) R Script

Console Terminal Background Jobs

R 4.3.1 - D:\r\prog\ ◊

neighbourhood_cleansed	neighbourhood_group_cleansed	latitude	longitude
Length:39453	Length:39453	Min. :40.50	Min. :-74.25
Class :character	Class :character	1st Qu.:40.69	1st Qu.:-73.98
Mode :character	Mode :character	Median :40.73	Median :-73.95
		Mean :40.73	Mean :-73.95
		3rd Qu.:40.76	3rd Qu.:-73.93
		Max. :40.91	Max. :-73.71

property_type	room_type	accommodates	bathrooms	bathrooms_text
Length:39453	Length:39453	Min. : 1.000	Mode:logical	Length:39453
Class :character	Class :character	1st Qu.: 2.000	NA's:39453	Class :character
Mode :character	Mode :character	Median : 2.000		Mode :character
		Mean : 2.934		
		3rd Qu.: 4.000		
		Max. :16.000		

bedrooms	beds	amenities	price	minimum_nights
Min. : 1.000	Min. : 1.000	Length:39453	Length:39453	Min. : 1.00
1st Qu.: 1.000	1st Qu.: 1.000	Class :character	Class :character	1st Qu.: 30.00
Median : 1.000	Median : 1.000	Mode :character	Mode :character	Median : 30.00
Mean : 1.597	Mean : 1.658			Mean : 28.05
3rd Qu.: 2.000	3rd Qu.: 2.000			3rd Qu.: 30.00
Max. :50.000	Max. :42.000			Max. :1250.00
NA's :16893	NA's :605			

maximum_nights	minimum_nights	maximum_nights	minimum_nights	minimum_maximum_nights	maximum_maximum_nights
Min. :1.000e+00	Min. : 1.00	Min. : 1.00	Min. : 1.00	Min. : 1.00e+00	Min. : 1.00e+00
1st Qu.:1.200e+02	1st Qu.: 30.00	1st Qu.: 30.00	1st Qu.: 30.00	1st Qu.:3.650e+02	1st Qu.:3.650e+02
Median :3.650e+02	Median : 30.00	Median : 30.00	Median : 30.00	Median :1.125e+03	Median :1.125e+03
Mean :1.606e+04	Mean : 29.63	Mean : 34.95	Mean : 6.549e+05	Mean : 1.222e+03	Mean : 1.222e+03
3rd Qu.:1.125e+03	3rd Qu.: 30.00	3rd Qu.: 30.00	3rd Qu.: 30.00	3rd Qu.:1.125e+03	3rd Qu.:1.125e+03
Max. :2.147e+09	Max. :1250.00	Max. :1250.00	Max. :1250.00	Max. :2.147e+09	Max. :2.147e+09

maximum_maximum_nights	minimum_nights_avg_ntm	maximum_nights_avg_ntm	calendar_updated
Min. :1.000e+00	Min. : 1.00	Min. :1.000e+00	Mode:logical
1st Qu.:3.650e+02	1st Qu.: 30.00	1st Qu.:3.650e+02	NA's:39453
Median :1.125e+03	Median : 30.00	Median :1.125e+03	
Mean :1.199e+06	Mean : 34.33	Mean :1.036e+06	
3rd Qu.:1.125e+03	3rd Qu.: 30.00	3rd Qu.:1.125e+03	
Max. :2.147e+09	Max. :1250.00	Max. :2.147e+09	

has_availability	availability_30	availability_60	availability_90	availability_365
Length:39453	Min. : 0.000	Min. : 0.00	Min. : 0.0	Min. : 0.0
Class :character	1st Qu.: 0.000	1st Qu.: 0.00	1st Qu.: 0.0	1st Qu.: 0.0
Mode :character	Median : 0.000	Median : 4.00	Median :16.0	Median :107.0
	Mean : 6.923	Mean :17.22	Mean :30.3	Mean :144.8
	3rd Qu.:11.000	3rd Qu.:34.00	3rd Qu.:61.0	3rd Qu.:291.0
	Max. :30.000	Max. :60.00	Max. :90.0	Max. :365.0

calendar_last_scraped	number_of_reviews	number_of_reviews_ltm	number_of_reviews_l30d
Length:39453	Min. : 0.00	Min. : 0.00	Min. : 0.0000
Class :character	1st Qu.: 0.00	1st Qu.: 0.00	1st Qu.: 0.0000
Mode :character	Median : 5.00	Median : 0.00	Median : 0.0000
	Mean : 25.84	Mean : 7.36	Mean : 0.6207
	3rd Qu.: 24.00	3rd Qu.: 6.00	3rd Qu.: 0.0000
	Max. :1834.00	Max. :686.00	Max. :97.0000

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Addins

proj.R*

```

14 tail(df)
15 colSums(is.na(df))
16 summary(df)
17 # mean(df),min(df),max(df),median(df) for dataset gives NA as most arguments are not numeric or log
18
19 library(dplyr)
20 df_no_duplicates_rows <- distinct(df)
21
22
23
24
25
26
27
28
29
30
31
32
33
42:1 (Top Level) :
```

Console Terminal < Background Jobs x

R 4.3.1 - D:/r prog/

```

review_scores_cleanliness review_scores_checkin review_scores_communication review_scores_location
Min. : 0.000 Min. : 0.000 Min. : 0.000 Min. : 0.000
1st Qu.: 4.500 1st Qu.: 4.800 1st Qu.: 4.800 1st Qu.: 4.620
Median : 4.800 Median : 4.940 Median : 4.960 Median : 4.830
Mean : 4.627 Mean : 4.813 Mean : 4.809 Mean : 4.725
3rd Qu.: 5.000 3rd Qu.: 5.000 3rd Qu.: 5.000 3rd Qu.: 5.000
Max. : 5.000 Max. : 5.000 Max. : 5.000 Max. : 5.000
NA's : 10644 NA's : 10658 NA's : 10650 NA's : 10661

review_scores_value license instant_bookable calculated_host_listings_count
Min. : 0.000 Mode:logical Length:39453 Min. : 1.00
1st Qu.: 4.520 NA's:39453 Class :character 1st Qu.: 1.00
Median : 4.750 Mode :character Median : 1.00
Mean : 4.627 Mean : 37.45
3rd Qu.: 4.930 3rd Qu.: 5.00
Max. : 5.000 Max. : 597.00
NA's : 10660

calculated_host_listings_count_entire_homes calculated_host_listings_count_private_rooms
Min. : 0.00 Min. : 0.00
1st Qu.: 0.00 1st Qu.: 0.00
Median : 1.00 Median : 0.00
Mean : 17.37 Mean : 19.98
3rd Qu.: 2.00 3rd Qu.: 2.00
Max. : 597.00 Max. : 519.00

calculated_host_listings_count_shared_rooms reviews_per_month
Min. : 0.00000 Min. : 0.010
1st Qu.: 0.00000 1st Qu.: 0.120
Median : 0.00000 Median : 0.450

```

Files Plots Packages Help Viewer Presentation

Import Dataset 161 MB

Environment History Connections Tutorial

Global Environment

- df 39453 obs. of 75 variables
- df_imputed 39453 obs. of 7 variables
- df_new 39453 obs. of 75 variables
- df_new1 39453 obs. of 75 variables
- df_no_duplicates_ 39453 obs. of 75 variables
- df_no_duplicates.. 39453 obs. of 75 variables
- df_no_na 0 obs. of 75 variables
- df1 28795 obs. of 75 variables
- neighbourhood fr 175 obs. of 2 variables

23°C Partly cloudy 22:49 22-11-2023 ENG IN

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Addins

proj.R*

```

10 Summary(ui)
11
12 # mean(df),min(df),max(df),median(df) for dataset gives NA as most arguments are not numeric or log
13
14 library(dplyr)
15 df_no_duplicates_rows <- distinct(df)
16 df_no_duplicates_columns <- df[, !duplicated(names(df))]
17 nrow(df)
18 ncol(df)
19 #indicates no complete duplicate rows or columns
20
21
22
23
24
25
26
27
28
29
30
31
32
33
42:1 (Top Level) :
```

Console Terminal < Background Jobs x

R 4.3.1 - D:/r prog/

```

calculated_host_listings_count_shared_rooms reviews_per_month
Min. : 0.00000 Min. : 0.010
1st Qu.: 0.00000 1st Qu.: 0.120
Median : 0.00000 Median : 0.450
Mean : 0.04111 Mean : 1.138
3rd Qu.: 0.00000 3rd Qu.: 1.620
Max. : 11.00000 Max. : 79.820
NA's : 10241

> library(dplyr)
> df_no_duplicates_rows <- distinct(df)
> df_no_duplicates_columns <- df[, !duplicated(names(df))]
> nrow(df)
[1] 39453
> ncol(df)
[1] 75
> #to print mean, median and std dev of some columns with int data type after removing duplicates
> #
> mean_of_host_listing_count <- mean(df$host_listings_count, na.rm = TRUE)
> median_of_host_listing_count <- median(df$host_listings_count, na.rm = TRUE)
> sd_of_host_listing_count <- sd(df$host_listings_count, na.rm = TRUE)

```

Files Plots Packages Help Viewer Presentation

Import Dataset 355 MB

Environment History Connections Tutorial

Global Environment

- mean_of_review_s... 4.81340059038027
- median_of_bedroo... 1
- median_of_host_l... 2
- median_of_host_t... 3
- median_of_review... 4.94
- numeric_vars Named logi [1:75] TRUE FALSE TRUE FALSE FALSE...
- s 'summaryDefault' Named num [1:6] 1 30 30 28.1...
- sd_of_bedrooms 0.976922305138487
- sd_of_host_listi... 620.086962405847
- sd_of_host_total 978.242964791906

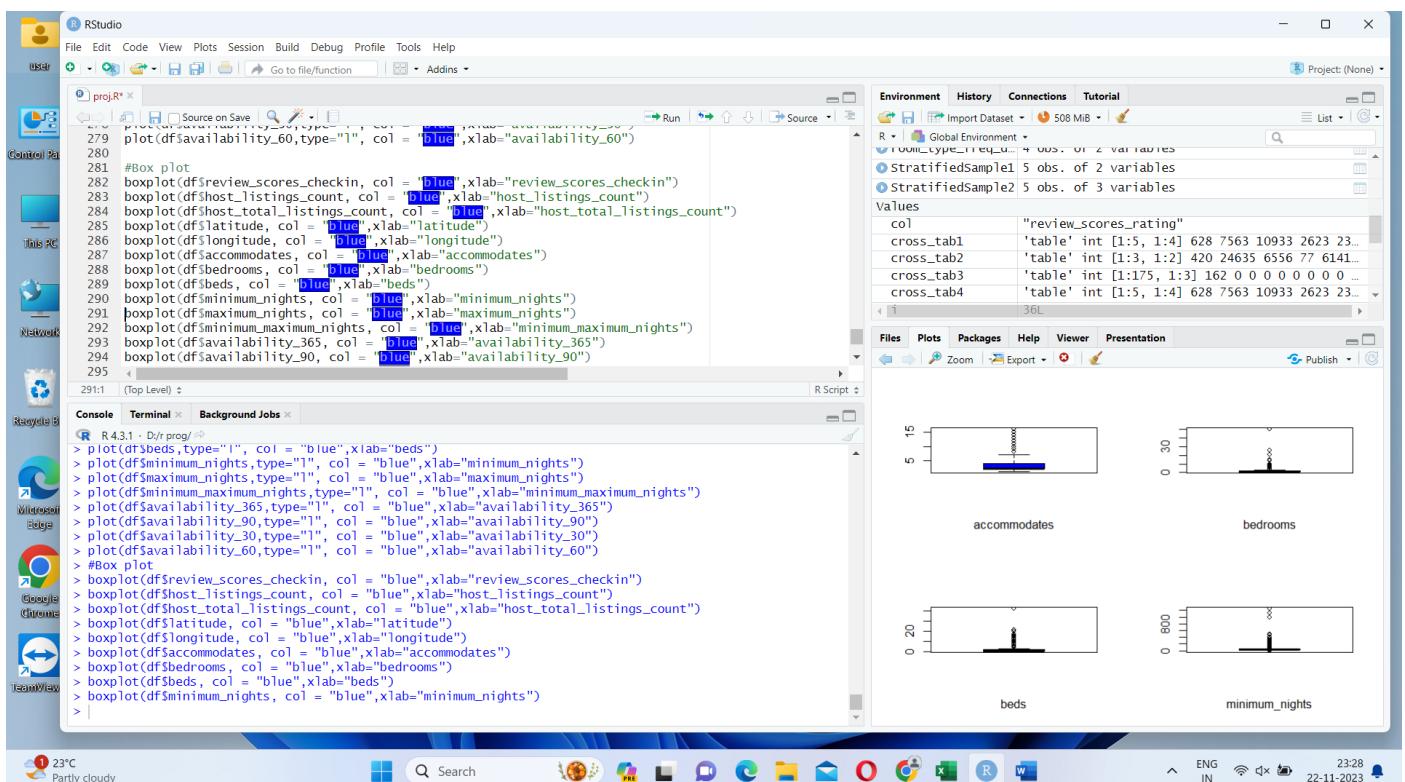
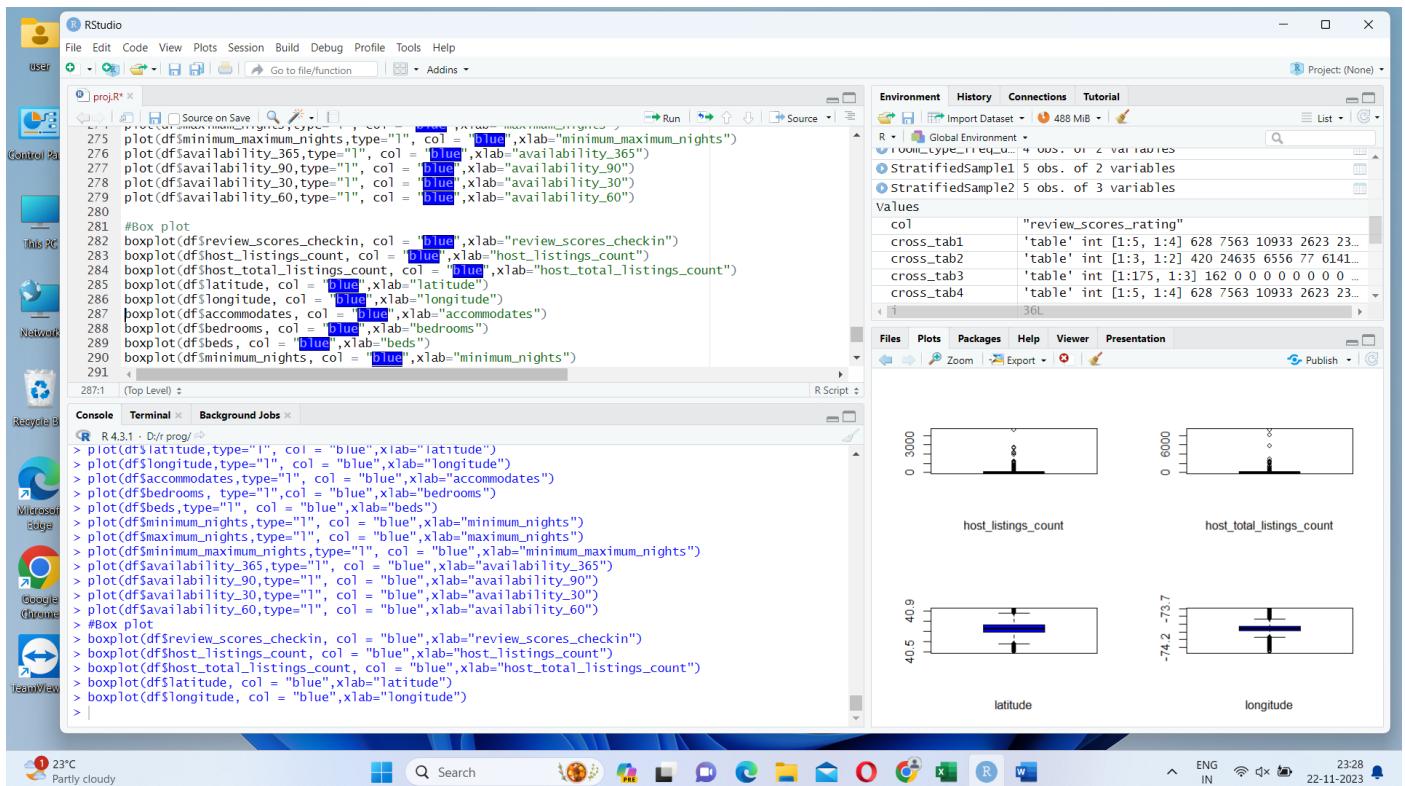
23°C Partly cloudy 22:52 22-11-2023 ENG IN

A screenshot of the RStudio interface. The left pane shows a code editor with R script code for generating boxplots. The right pane displays the resulting boxplot titled "Boxplot of Review Scores by Neighbourhood Group". The x-axis is labeled "neighbourhood_group_cleaned" and has categories for Bronx, Brooklyn, Manhattan, Queens, and Staten Island. The y-axis is labeled "review_scores_rating" and ranges from 0 to 5. Each category has a black box representing the interquartile range, a horizontal line inside the box for the median, and vertical whiskers extending to the minimum and maximum values. Numerous individual data points (outliers) are plotted as small black dots.

The screenshot shows an RStudio interface with the following details:

- File Menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Toolbar:** Source on Save, Run, Source.
- Code Editor:** A large pane containing R code for plotting various metrics like bedrooms, nights, and availability across different time periods (90, 30, and 60 days). The code uses blue text for comments and labels.
- Environment Tab:** Shows the global environment with objects like `df$accommodates`, `df$bedrooms`, etc., and four cross-tabulation tables (`cross_tab1` to `cross_tab4`) with their respective sizes.
- Plots:** Four box plots are displayed in the main area:
 - `df$availability_90`: Availability over 90 days, showing a median around 20,000.
 - `df$availability_30`: Availability over 30 days, showing a median around 10,000.
 - `df$availability_60`: Availability over 60 days, showing a median around 10,000.
 - `review_scores_checkin`: Review scores for check-in, showing a median around 4.5.
- Console Tab:** Displays the R code being run, identical to the code in the editor.
- System Status Bar:** Shows the date (22-11-2023), time (23:27), and weather (23°C, Partly cloudy).

Last one in this image



3) The coorelation matrix for the columns host_listings_count , host_total_listings_count , accommodates, bathrooms,bedrooms ,beds, review_scores_rating was tabulated

Bedrooms and review score rating showed negative correlation with host_listings_count

Review score rating also showed negative correlation with host_total_listings_count

All other pairs showed a positive correlation

Negative correlation indicates that as one variable decreases the other one decreases too and vice versa for positive
Thus it is seen that as host listings count decreased so did the bed room and review score rating decrease and when host total listing count decreased the review score rating also decreased.

The screenshot shows an RStudio interface. In the top-left, there's a sidebar with icons for Control Panel, This PC, Network, and Recycle Bin. The main area has tabs for 'proj.R' (active), 'Source on Save', 'Run', 'Source', 'Console', 'Terminal', and 'Background Jobs'. The 'Console' tab shows R code for imputing missing values and calculating correlation matrices. The 'Environment' tab lists various data frames like df, df_imputed, df_new, etc. The 'Plots' tab is active, showing a scatter plot of cor_matrix_imputed. The status bar at the bottom indicates it's 23:10, ENG IN, and the date is 22-11-2023.

```

124 # Impute missing values with mean
125 df_imputed <- df[, c("host_listings_count", "host_total_listings_count", "accommodates", "bathrooms",
126
127 ~ for (col in colnames(df_imputed)) {
128   df_imputed[, col] <- ifelse(is.na(df_imputed[, col]), mean(df_imputed[, col], na.rm = TRUE), df_i
129 }
130
131 cor_matrix_imputed <- cor(df_imputed)
132 cor_matrix_imputed
133
134
135 df

```

	host_listings_count	host_total_listings_count	accommodates	bathrooms
host_listings_count	1.000000000	0.914563491	0.05786284	NA
host_total_listings_count	0.914563491	1.000000000	0.07586933	NA
accommodates	0.057862842	0.075869326	1.00000000	NA
bathrooms	NA	NA	NA	1
bedrooms	-0.009693927	0.005757229	0.53704444	NA
beds	0.013513842	0.031557376	0.76804195	NA
review_scores_rating	-0.04555538	-0.013450803	0.02229621	NA
	bedrooms	beds	review_scores_rating	
host_listings_count	-0.009693927	0.01351384	-0.01455554	
host_total_listings_count	0.005757229	0.031557378	-0.01345080	
accommodates	0.537044440	0.76804195	0.02229621	
bathrooms	NA	NA	NA	
bedrooms	1.000000000	0.61682916	0.01293431	
beds	0.616829160	1.00000000	0.02682721	
review_scores_rating	0.012934306	0.02682721	1.00000000	

4) When frequency distributions for some variables were tabulated we saw that New York ,United States had highest count for neighbourhood.

```

101           New York city, New York, United States      1
102           New York, New York , United States        1
103           New York, Ny, United States               1
104           New York, US, New York, United States     1
105           New York, United States                 9780
106           New york, New York, United States       2
107           New-York, New York, United States        1
108           Newyork, New York, United States        1
109           Nueva York, New York, United States     1

```

Entire rental unit had the highest property type listing

```

11           Entire guest suite    340
12           Entire guesthouse    66
13           Entire home          1577
14           Entire home/apt     11
15           Entire loft          596
16           Entire place         83
17           Entire rental unit  16607
18           Entire serviced apartment 673
19           Entire townhouse     537
20           Entire vacation home 40
21           Entire villa          7
22           Farm stay             1

```

Entire home/apt had the highest room type and hotel room had least in dataset

The screenshot shows the RStudio interface. The left pane displays an R script named 'projR.R' with several code examples. The right pane shows the 'Environment' tab with a list of objects and their details. The bottom status bar shows system information like weather, date, and time.

```
160 # Example 3: Cross-tabulation for neighbourhood and host_is_superhost
161 cross_tab3 <- table(df$neighbourhood, df$host_is_superhost)
162 cross_tab3
163
164 # Example 4: Cross-tabulation for neighbourhood_group_cleansed and room_type
165 cross_tab4 <- table(df$neighbourhood_group_cleansed, df$room_type)
166 cross_tab4
167 cross_tab4
168
169
153:1 (Top Level) ▾
```

R 4.3.1 · D:/r/proj/ ↵

room_type	n
Entire home/apt	21981
Hotel room	133
Private room	16841
Shared room	498

> # Example 3: Frequency distribution for room_type
> room_type_freq_dplyr <- df %>% count(room_type)
> room_type_freq_dplyr

room_type	n
Entire home/apt	21981
Hotel room	133
Private room	16841
Shared room	498

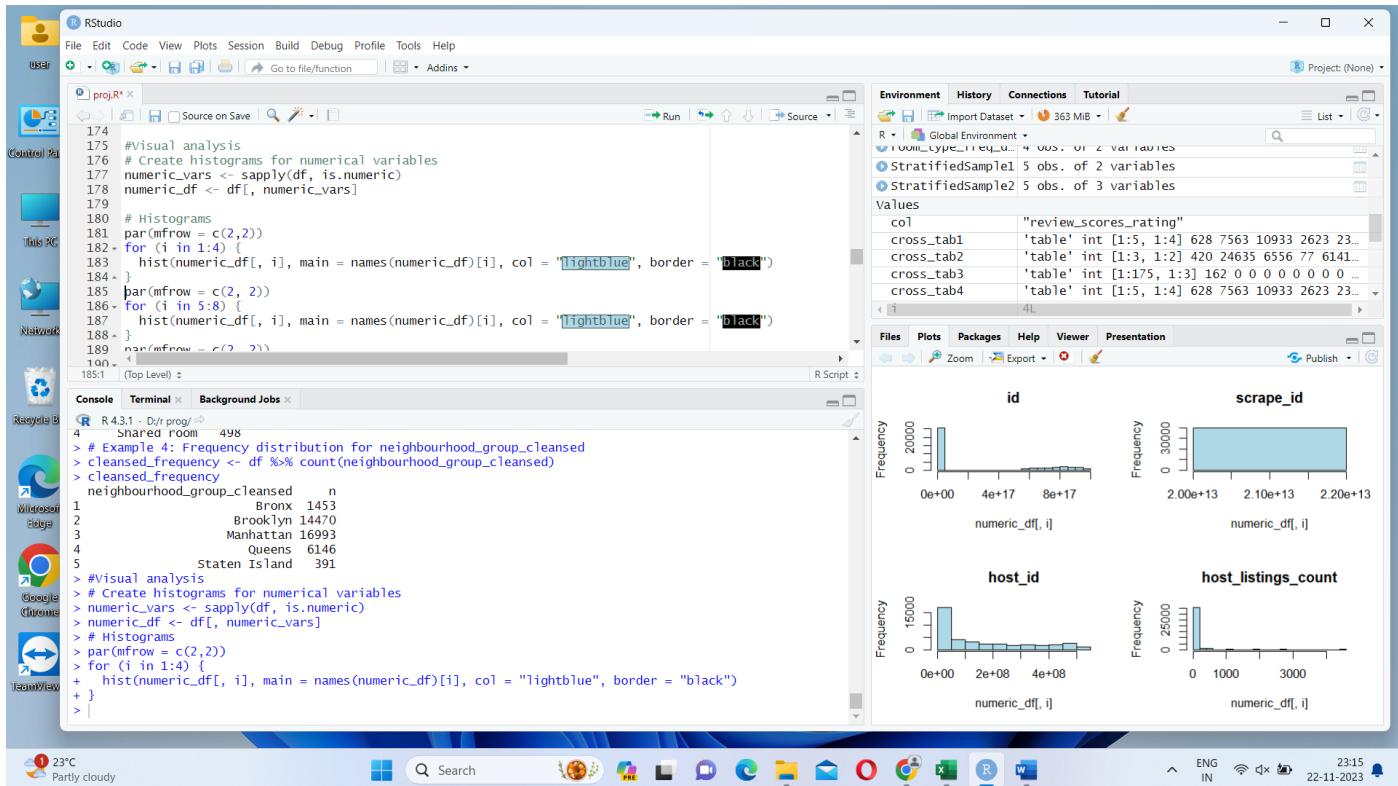
> # Example 4: Frequency distribution for neighbourhood_group_cleansed
> cleansed_frequency <- df %>% count(neighbourhood_group_cleansed)
> cleansed_frequency

neighbourhood_group_cleansed	n
Bronx	1453
Brooklyn	14470
Manhattan	16993
Queens	6146
Staten Island	391

Manhattan had the highest neighbourhood_group_cleansed listing and Staten island had the least in the dataset

5) From the histograms and density plots :

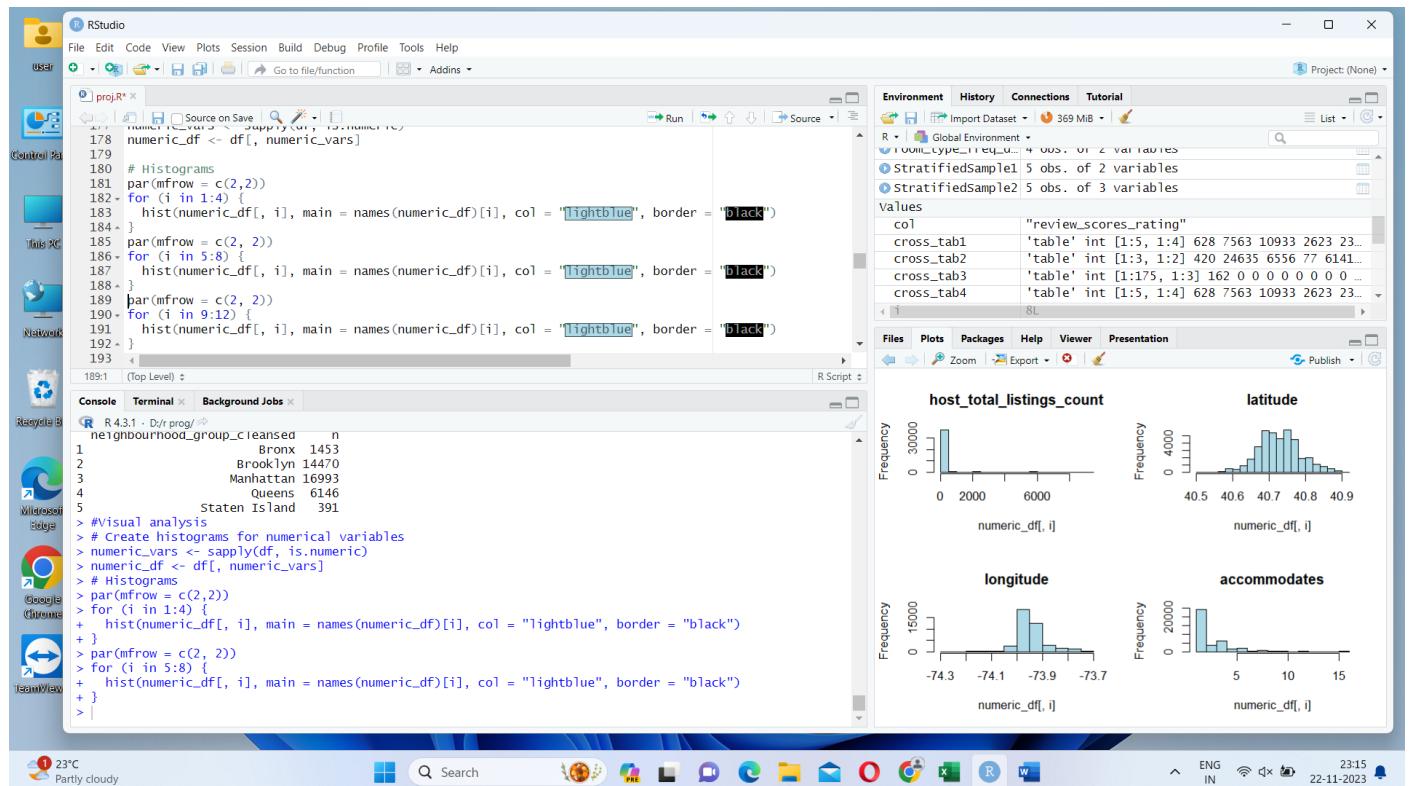
i) Most host listings are less than 1000



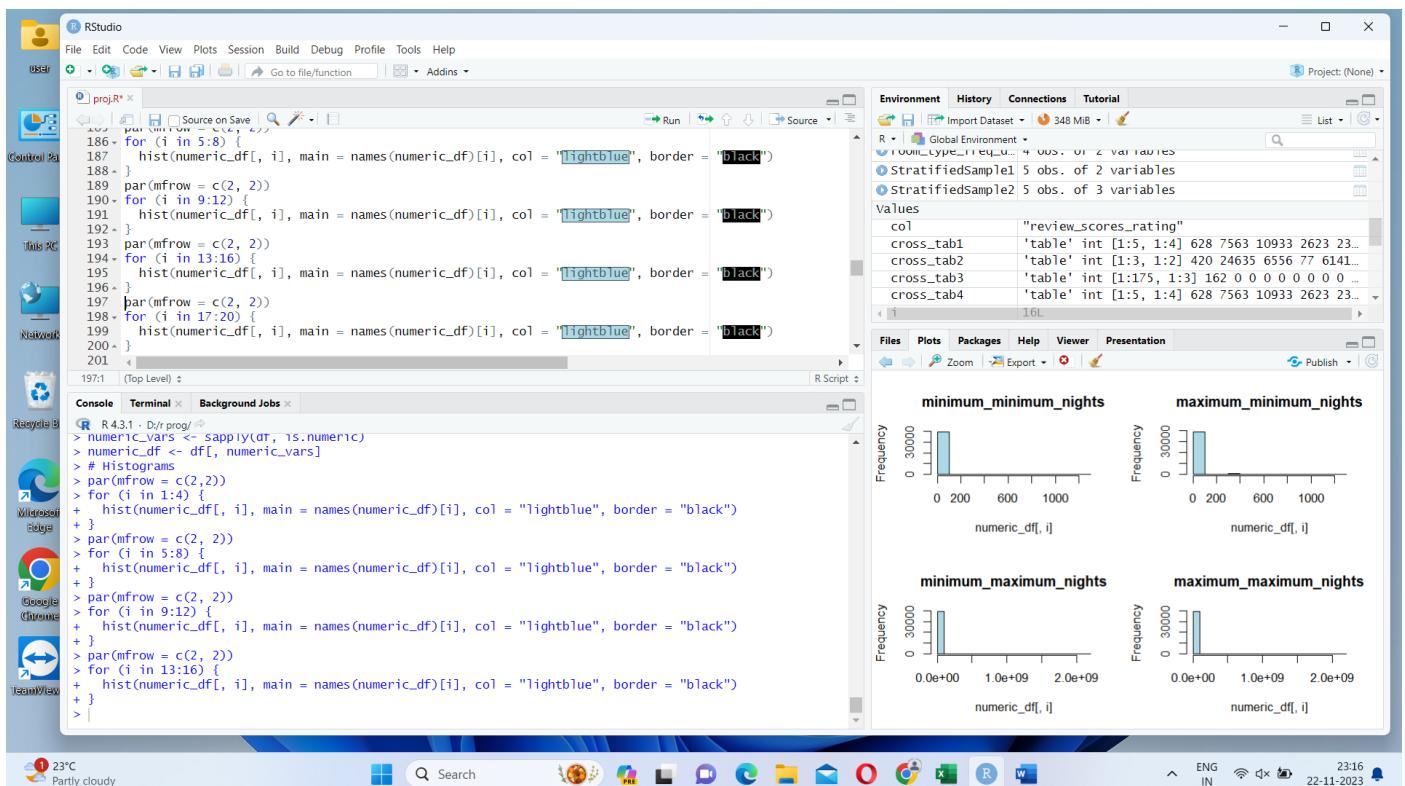
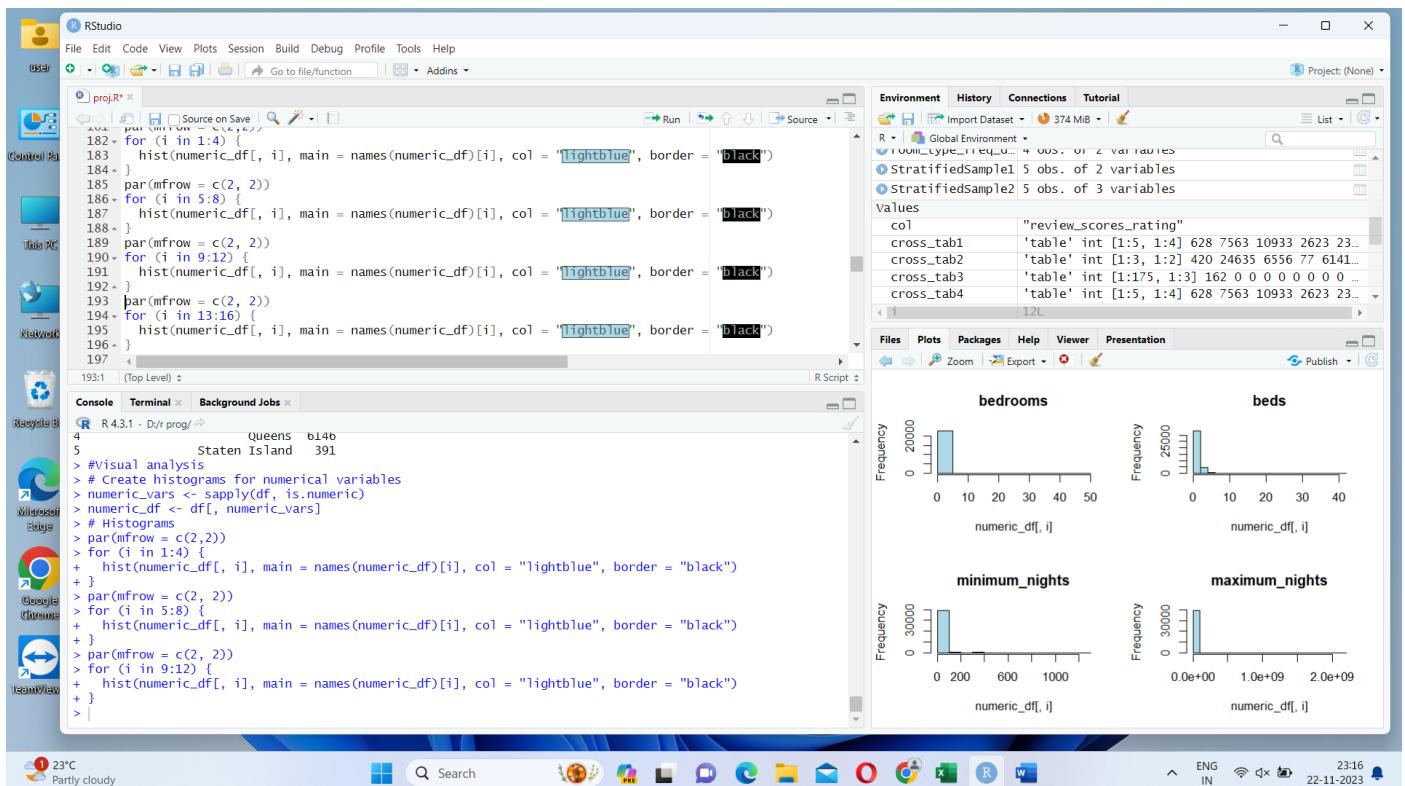
ii) Most host total listings are less than 1000

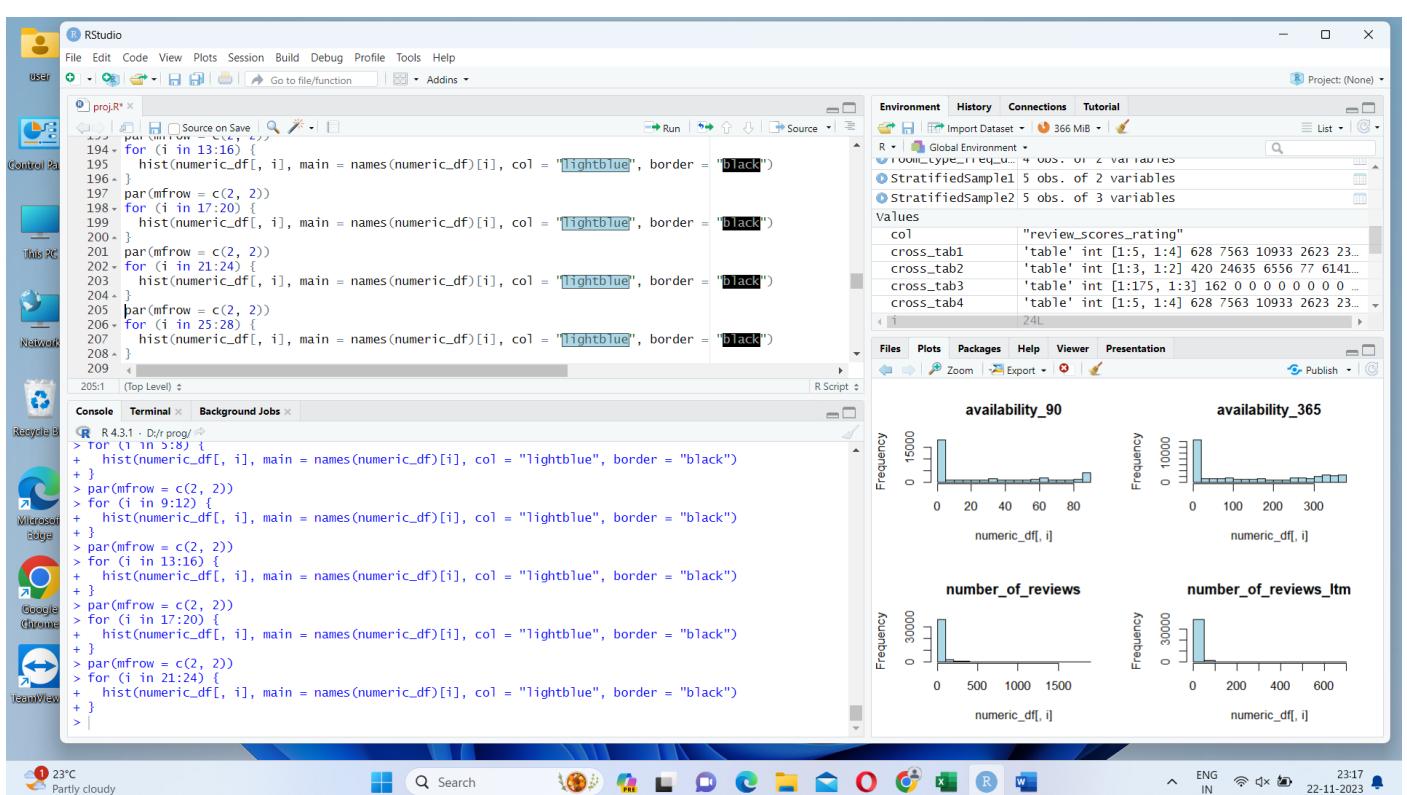
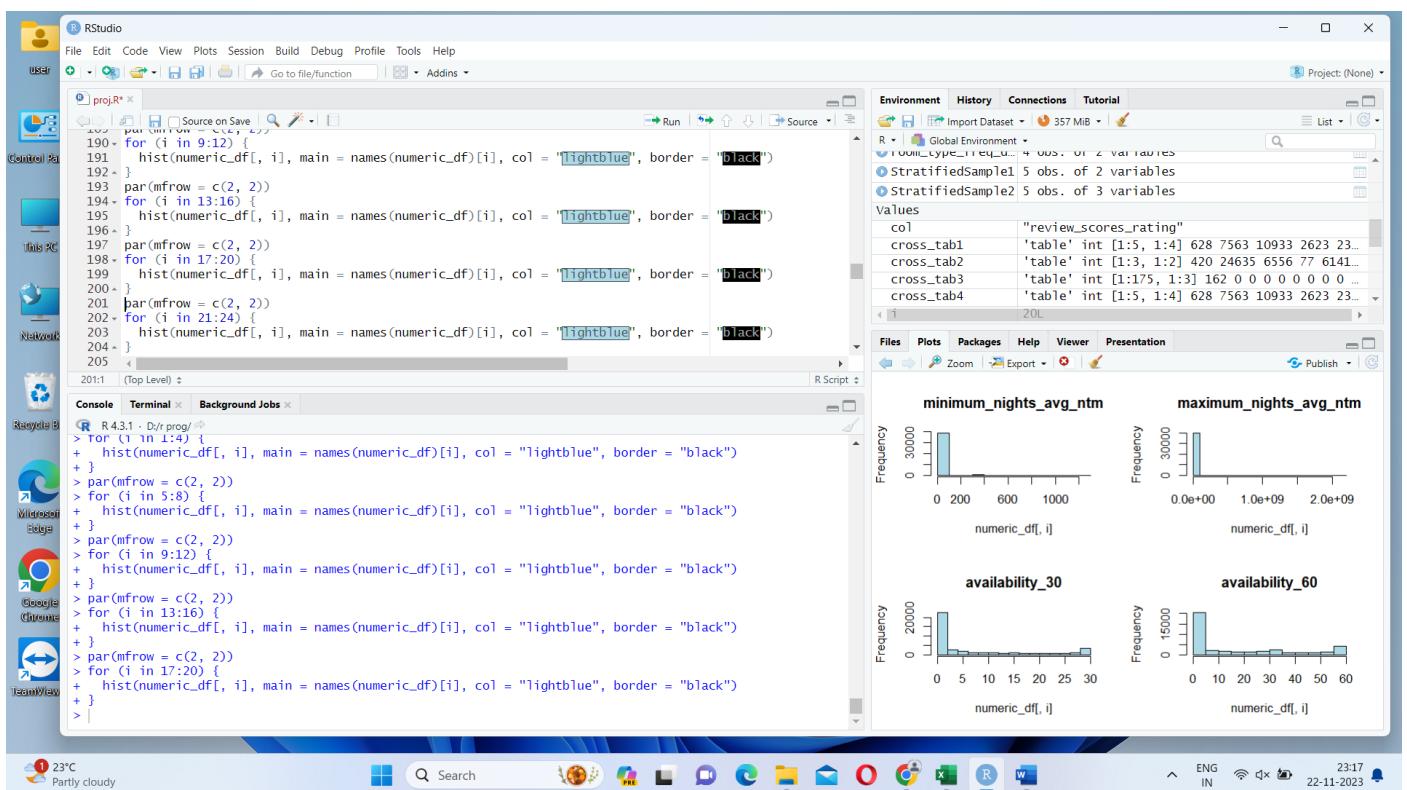
iii) Latitude values show a normal distribution

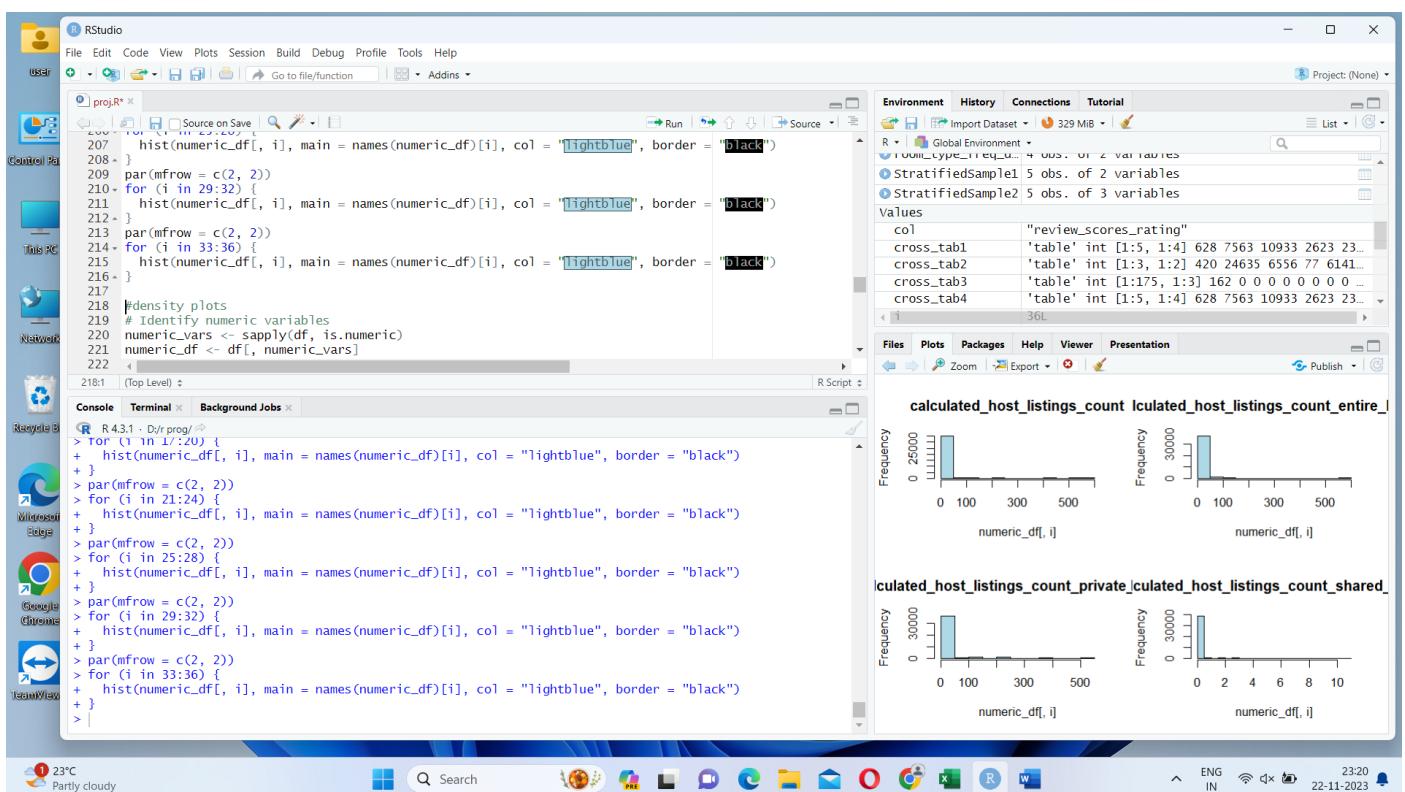
iv) Most listings have accommodates less than 5



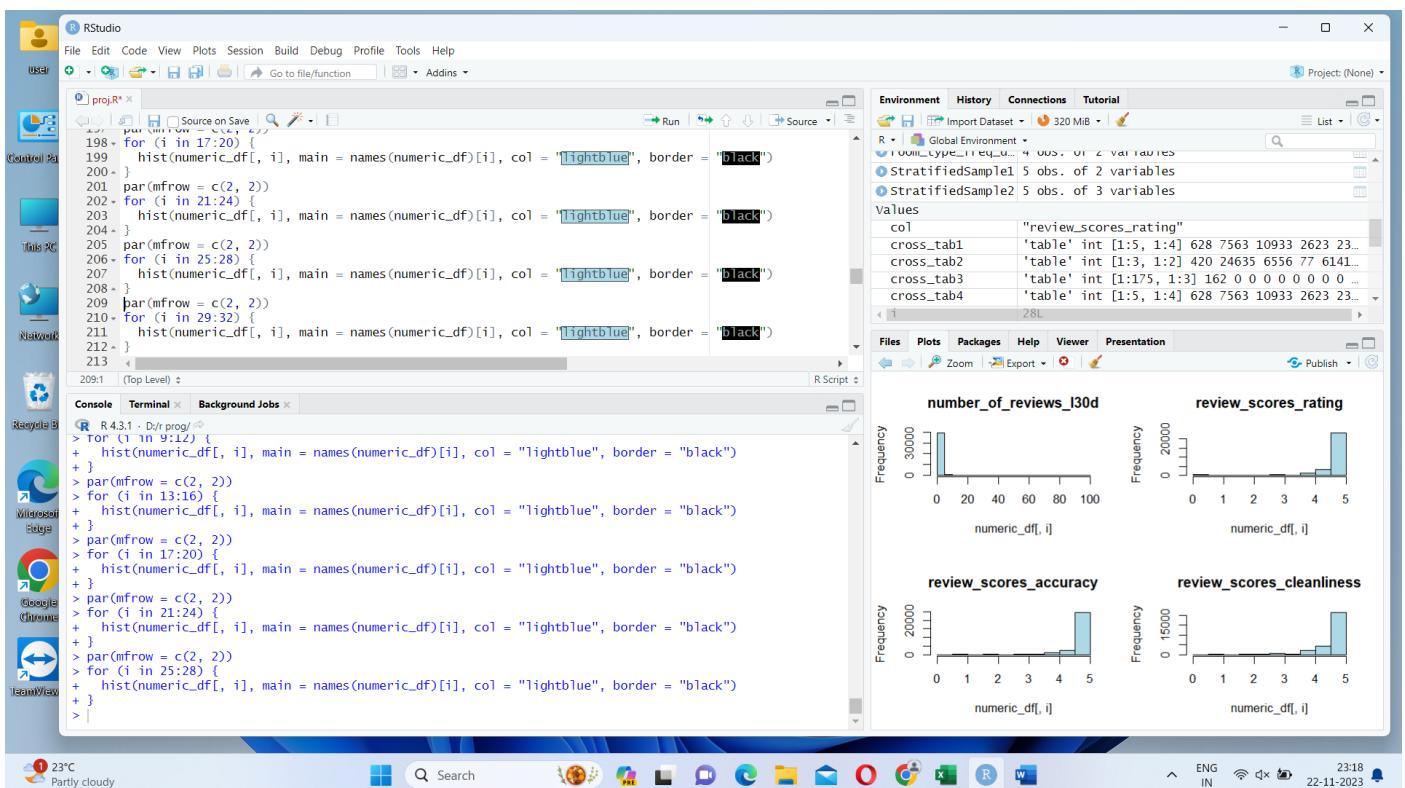
v) Most categorical variable values are all concentrated below a particular value(right skewed (right tail)plot)

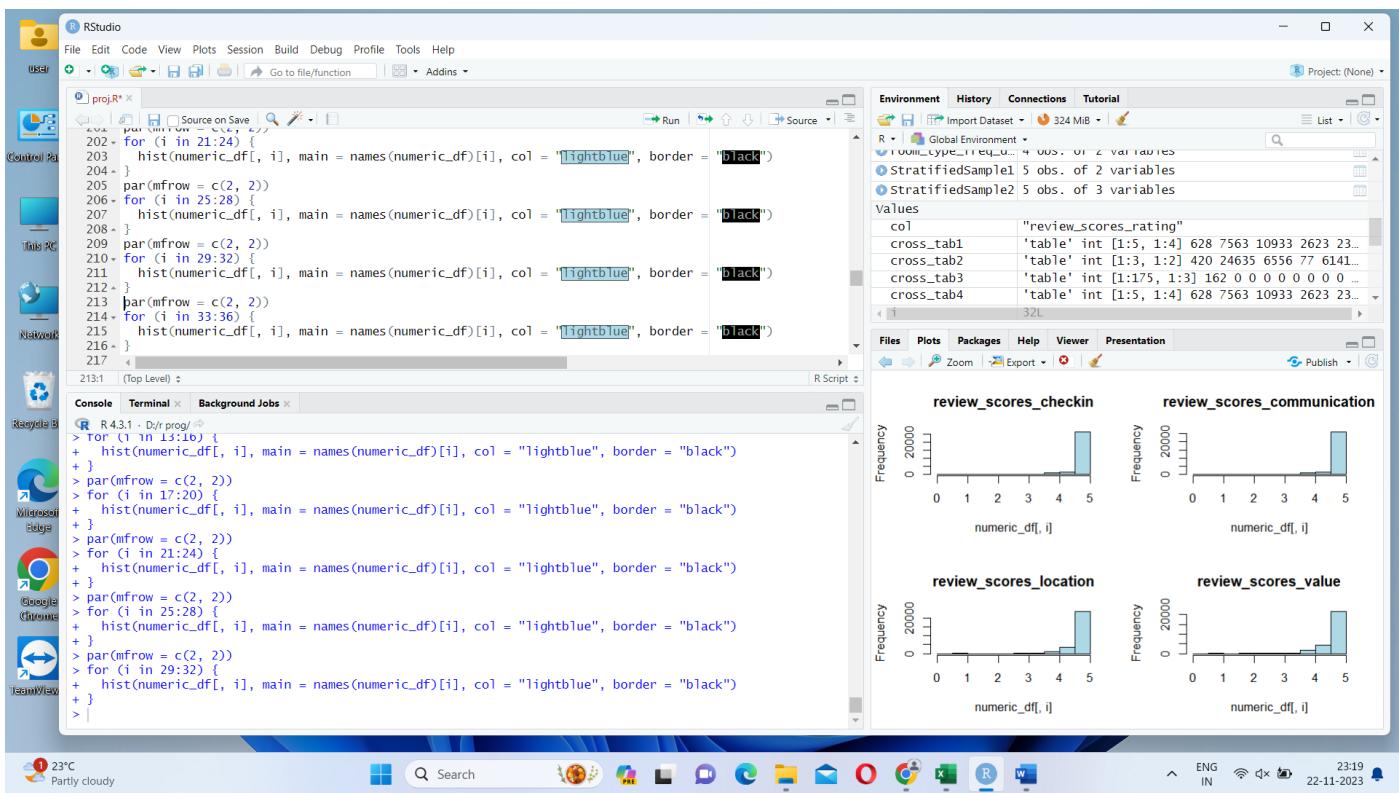




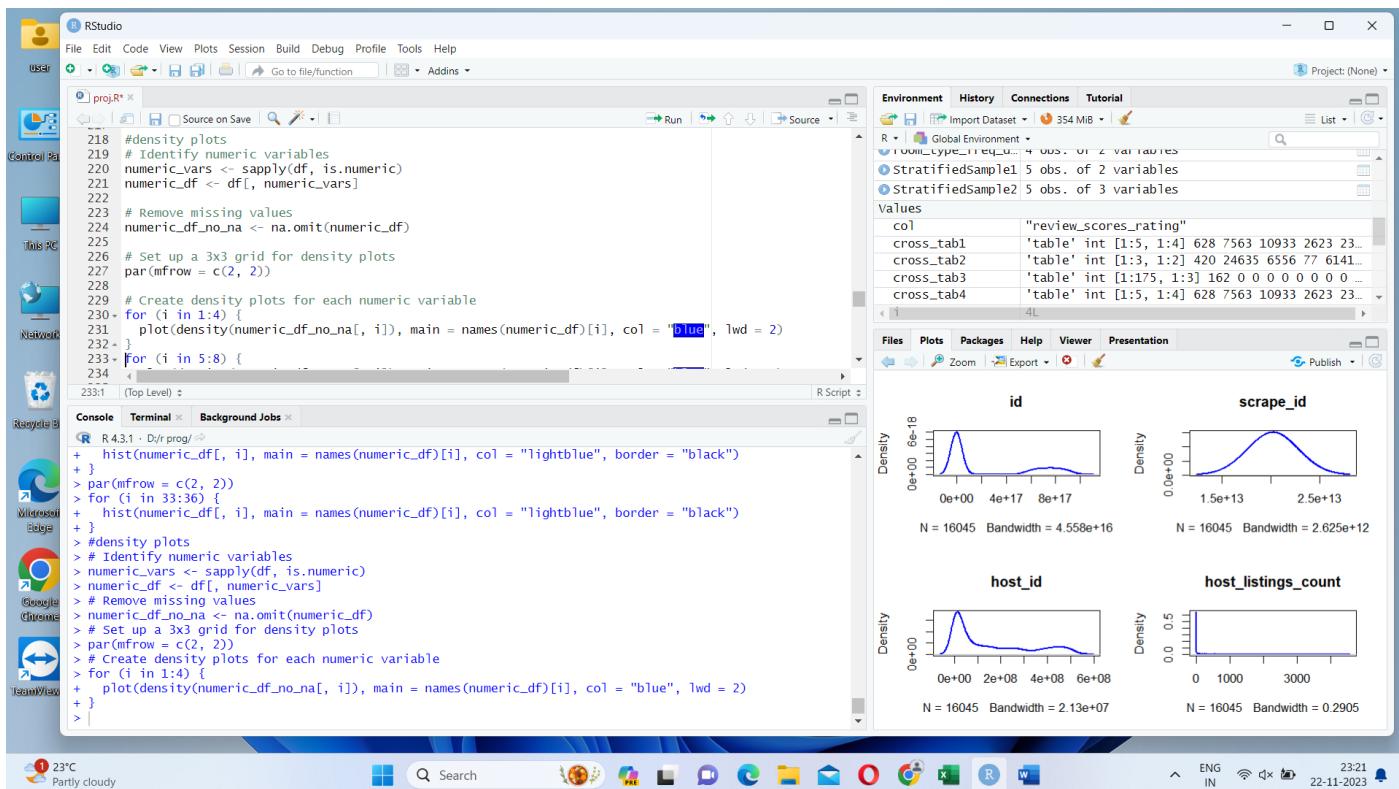


vi) all review related columns have most values above a particular value(left skewed plot)

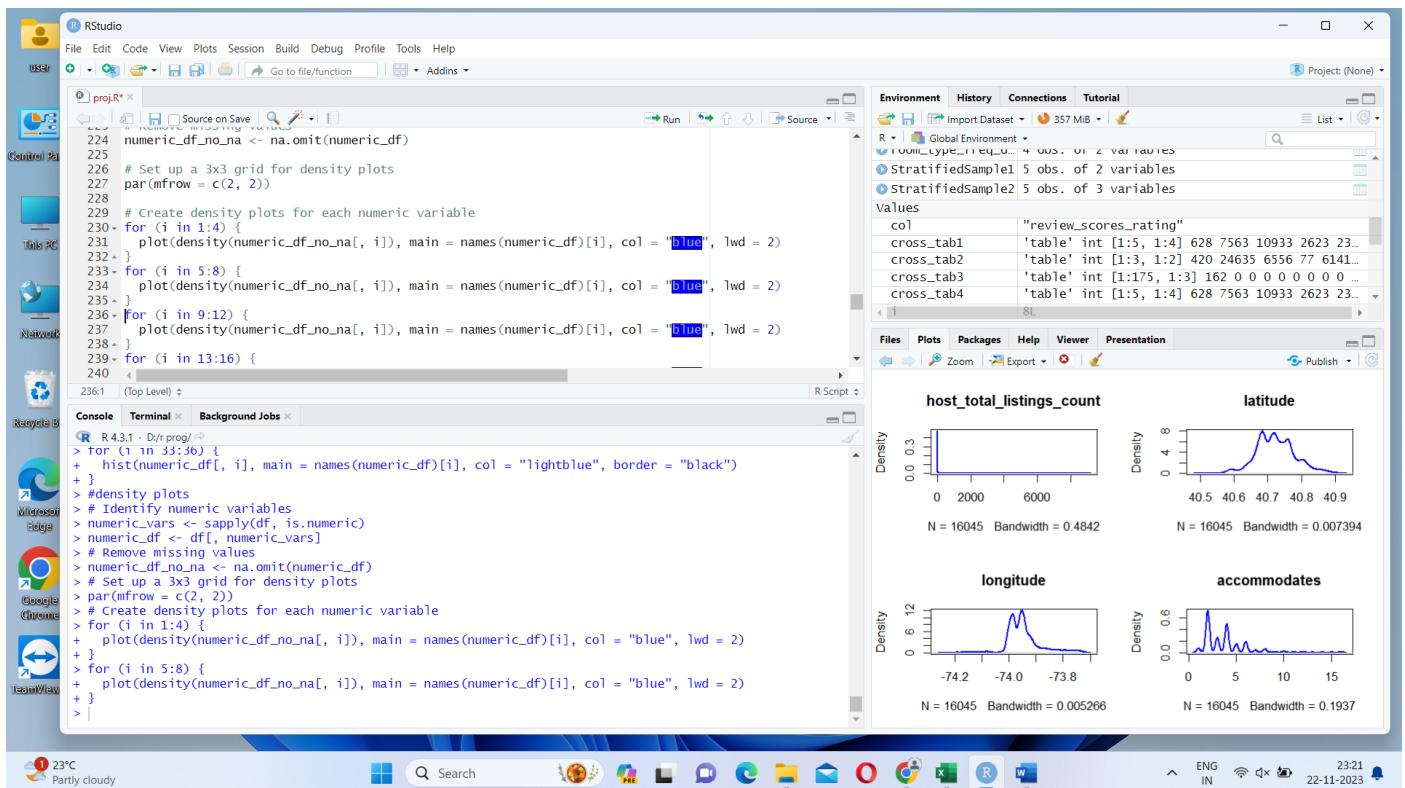




vii) Host id and id show right skewed while scrape id is normally distributed



viii) Longitude is bimodal(2 peaks) and latitude is almost normal with 3 peaks near the highest point of the curve



the density plots too show the same inference screenshots are given above from page number 26 to 30

6) There are no duplicates (rows/columns) seen in this dataset

The screenshot shows the RStudio interface with the following details:

- File Menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Project Bar:** Shows 'User' and 'Control Panel'.
- Console Tab:** Displays R version 4.3.1 and the path D:/r/prog/. It shows the execution of R code to remove duplicates from a dataset and prints the summary statistics for the 'host_listings_count' column.
- Environment Tab:** Shows the global environment with variables like mean_of_review, median_of_bedrooms, median_of_host, median_of_review, numeric_vars, s, sd_of_bedrooms, sd_of_host_listings, and sd_of_host_total.
- Code Editor:** A script named 'proj.R' containing R code to remove duplicates and calculate summary statistics.
- System Taskbar:** Shows weather (23°C, Partly cloudy), search bar, and various system icons.

```
library(dplyr)
df_no_duplicates_rows <- distinct(df)
df_no_duplicates_columns <- df[, !duplicated(names(df))]
nrow(df)
ncol(df)
#indicates no complete duplicate rows or columns

#to print mean, median and std dev of some columns with int data type after removing duplicates
#1
mean_of_host_listing_count <- mean(df$host_listings_count, na.rm = TRUE)
median_of_host_listing_count <- median(df$host_listings_count, na.rm = TRUE)
sd_of_host_listing_count <- sd(df$host_listings_count, na.rm = TRUE)
```

calculated_host_listings_count_shared_rooms	reviews_per_month
Min. : 0.00000	Min. : 0.010
1st Qu.: 0.00000	1st Qu.: 0.120
Median : 0.00000	Median : 0.450
Mean : 0.04111	Mean : 1.138
3rd Qu.: 0.00000	3rd Qu.: 1.620
Max. : 11.00000	Max. : 79.820
	NA's : 10241