

W3 In-class practice

Jadon Fowler jaf582 5778191

2022-09-15

Instructions

Download a copy of this markdown. Change the **author:** tag above to have your name and NAU's ID. Fill in the chunks with your R code that solves the problem. Knit your final document into a HTML file and submit it through BBLearn in the assignment **[In-class]** CSV before the end of the class.

If you're using any libraries to run your code, please load them here.

Problem 1

Download the .csv file attached to this assignment. Write an R code that imports the CVS into R. It's your job to define whether you need or not to add parameters to the function you'll use to import the data. The variable used to store the csv must be named **airbnb**. Print the **head** and the **tail** of the dataset.

```
#write your R code here
airbnb <- read.csv("Airbnb_Open_Data.csv")
head(airbnb) #lookup the head
```

```
##           id                                NAME      host.id
## 1 1001254          Clean & quiet apt home by the park 80014485718
## 2 1002102                                Skylit Midtown Castle 52335172823
## 3 1002403          THE VILLAGE OF HARLEM....NEW YORK ! 78829239556
## 4 1002755                                           85098326012
## 5 1003689 Entire Apt: Spacious Studio/Loft by central park 92037596077
## 6 1004098          Large Cozy 1 BR Apartment In Midtown East 45498551794
##  host_identity_verified host.name neighbourhood.group neighbourhood      lat
## 1             unconfirmed Madaline           Brooklyn  Kensington 40.64749
## 2              verified   Jenna           Manhattan  Midtown 40.75362
## 3                   Elise           Manhattan    Harlem 40.80902
## 4             unconfirmed   Garry           Brooklyn Clinton Hill 40.68514
## 5              verified   Lyndon           Manhattan East Harlem 40.79851
## 6              verified Michelle           Manhattan Murray Hill 40.74767
##           long      country country.code instant_bookable cancellation_policy
## 1 -73.97237 United States      US          FALSE          strict
## 2 -73.98377 United States      US          FALSE          moderate
## 3 -73.94190 United States      US           TRUE          flexible
## 4 -73.95976 United States      US           TRUE          moderate
## 5 -73.94399 United States      US          FALSE          moderate
## 6 -73.97500 United States      US           TRUE          flexible
##           room.type Construction.year price service.fee minimum.nights
```

```

## 1 Private room 2020 $966 $193 10
## 2 Entire home/apt 2007 $142 $28 30
## 3 Private room 2005 $620 $124 3
## 4 Entire home/apt 2005 $368 $74 30
## 5 Entire home/apt 2009 $204 $41 10
## 6 Entire home/apt 2013 $577 $115 3
## number.of.reviews last.review reviews.per.month review.rate.number
## 1 9 10/19/2021 0.21 4
## 2 45 5/21/2022 0.38 4
## 3 0 NA 5
## 4 270 7/5/2019 4.64 4
## 5 9 11/19/2018 0.10 3
## 6 74 6/22/2019 0.59 3
## calculated.host.listings.count availability.365
## 1 6 286
## 2 2 228
## 3 1 352
## 4 1 322
## 5 1 289
## 6 1 374
##
## 1
## 2 Pet friendly but please confirm with me if the pet you are planning on bringing with you is OK. I
## 3 I encourage you to use my kitchen, cooking and laundry facilities. Th
## 4
## 5
## 6
## license
## 1
## 2
## 3
## 4
## 5
## 6

```

```
tail(airbnb) #lookup the tail
```

```

## id NAME host.id
## 102594 6091885 Welcoming, Clean, Cheap on St Marks 33188605074
## 102595 6092437 Spare room in Williamsburg 12312296767
## 102596 6092990 Best Location near Columbia U 77864383453
## 102597 6093542 Comfy, bright room in Brooklyn 69050334417
## 102598 6094094 Big Studio-One Stop from Midtown 11160591270
## 102599 6094647 585 sf Luxury Studio 68170633372
## host_identity_verified host.name neighbourhood.group
## 102594 verified Felipe Manhattan
## 102595 verified Krik Brooklyn
## 102596 unconfirmed Mifan Manhattan
## 102597 unconfirmed Megan Brooklyn
## 102598 unconfirmed Christopher Queens
## 102599 unconfirmed Rebecca Manhattan
## neighbourhood lat long country country.code
## 102594 East Village 40.72826 -73.98422 United States US
## 102595 Williamsburg 40.70862 -73.94651 United States US

```

```

## 102596 Morningside Heights 40.80460 -73.96545 United States US
## 102597 Park Slope 40.67505 -73.98045 United States US
## 102598 Long Island City 40.74989 -73.93777 United States US
## 102599 Upper West Side 40.76807 -73.98342 United States US
## instant_bookable cancellation_policy room.type Construction.year
## 102594 TRUE strict Private room 2017
## 102595 FALSE flexible Private room 2003
## 102596 TRUE moderate Private room 2016
## 102597 TRUE moderate Private room 2009
## 102598 TRUE strict Entire home/apt 2015
## 102599 FALSE flexible Entire home/apt 2010
## price service.fee minimum.nights number.of.reviews last.review
## 102594 $1,099 $220 1 8 9/6/2015
## 102595 $844 $169 1 0
## 102596 $837 $167 1 1 7/6/2015
## 102597 $988 $198 3 0
## 102598 $546 $109 2 5 10/11/2015
## 102599 $1,032 $206 1 0
## reviews.per.month review.rate.number calculated.host.listings.count
## 102594 0.16 4 2
## 102595 NA 3 1
## 102596 0.02 2 2
## 102597 NA 5 1
## 102598 0.10 3 1
## 102599 NA 3 1
## availability.365
## 102594 152
## 102595 227
## 102596 395
## 102597 342
## 102598 386
## 102599 69
##
## 102594
## 102595
## 102596 House rules: Guests agree to the following terms and conditions 1.Guest(s) agree to NO PARTIES
## 102597
## 102598
## 102599
## license
## 102594
## 102595
## 102596
## 102597
## 102598
## 102599

```

Problem 2

Check the following information about the dataset:

1. the structure

```
#write your R code here
str(airbnb)
```

```
## 'data.frame': 102599 obs. of 26 variables:
## $ id : int 1001254 1002102 1002403 1002755 1003689 1004098 1004650 1005...
## $ NAME : chr "Clean & quiet apt home by the park" "Skylit Midtown Castle"
## $ host.id : num 8.00e+10 5.23e+10 7.88e+10 8.51e+10 9.20e+10 ...
## $ host_identity_verified : chr "unconfirmed" "verified" "" "unconfirmed" ...
## $ host.name : chr "Madaline" "Jenna" "Elise" "Garry" ...
## $ neighbourhood.group : chr "Brooklyn" "Manhattan" "Manhattan" "Brooklyn" ...
## $ neighbourhood : chr "Kensington" "Midtown" "Harlem" "Clinton Hill" ...
## $ lat : num 40.6 40.8 40.8 40.7 40.8 ...
## $ long : num -74 -74 -73.9 -74 -73.9 ...
## $ country : chr "United States" "United States" "United States" "United States"
## $ country.code : chr "US" "US" "US" "US" ...
## $ instant_bookable : logi FALSE FALSE TRUE TRUE FALSE TRUE ...
## $ cancellation_policy : chr "strict" "moderate" "flexible" "moderate" ...
## $ room.type : chr "Private room" "Entire home/apt" "Private room" "Entire home"
## $ Construction.year : int 2020 2007 2005 2005 2009 2013 2015 2009 2005 2015 ...
## $ price : chr "$966 " "$142 " "$620 " "$368 " ...
## $ service.fee : chr "$193 " "$28 " "$124 " "$74 " ...
## $ minimum.nights : int 10 30 3 30 10 3 45 45 2 2 ...
## $ number.of.reviews : int 9 45 0 270 9 74 49 49 430 118 ...
## $ last.review : chr "10/19/2021" "5/21/2022" "" "7/5/2019" ...
## $ reviews.per.month : num 0.21 0.38 NA 4.64 0.1 0.59 0.4 0.4 3.47 0.99 ...
## $ review.rate.number : int 4 4 5 4 3 3 5 5 3 5 ...
## $ calculated.host.listings.count : int 6 2 1 1 1 1 1 1 1 1 ...
## $ availability.365 : int 286 228 352 322 289 374 224 219 180 375 ...
## $ house_rules : chr "Clean up and treat the home the way you'd like your home to
## $ license : chr "" "" "" "" ...
```

2. the dimension

```
#write your R code here
dim(airbnb)
```

```
## [1] 102599 26
```

3. the column names

```
#write your R code here
colnames(airbnb)
```

```
## [1] "id" "NAME"
## [3] "host.id" "host_identity_verified"
## [5] "host.name" "neighbourhood.group"
## [7] "neighbourhood" "lat"
## [9] "long" "country"
## [11] "country.code" "instant_bookable"
## [13] "cancellation_policy" "room.type"
## [15] "Construction.year" "price"
```

```
## [17] "service.fee"           "minimum.nights"
## [19] "number.of.reviews"    "last.review"
## [21] "reviews.per.month"    "review.rate.number"
## [23] "calculated.host.listings.count" "availability.365"
## [25] "house_rules"          "license"
```

Problem 3

It looks like the column names are quite inconvenient mainly because they have white space between words, some are pretty long, and they do not follow any standard format (lower vs. upper case, underscore as delimiters, and so on).

Considering that, change the column names to a more convenient, standardized format. Below you'll find my solution, but you can create your own!

```
#write your R code here
# makes everything lowercase and replaces '_' with '.'
airbnb <- rename_with(airbnb, function(x) tolower(str_replace_all(x, "_", ".")))
colnames(airbnb)
```

```
## [1] "id"           "name"
## [3] "host.id"      "host.identity.verified"
## [5] "host.name"    "neighbourhood.group"
## [7] "neighbourhood" "lat"
## [9] "long"         "country"
## [11] "country.code" "instant.bookable"
## [13] "cancellation.policy" "room.type"
## [15] "construction.year" "price"
## [17] "service.fee"     "minimum.nights"
## [19] "number.of.reviews" "last.review"
## [21] "reviews.per.month" "review.rate.number"
## [23] "calculated.host.listings.count" "availability.365"
## [25] "house.rules"      "license"
```

Problem 4

When we inspect the structure, we see that there are a few variables that has the incorrect type. For example, the columns `price` and `service fee` (you may have changed these names!) has the dollar sign at the begining, which turns it into character instead of numeric values.

To fix that problem, write a **regex** that detects and remove the dollar sign from those two columns.

Note: you may want to attribute the column you're going to work with to a variable, work with that array, and then replace the column with the clean array. I started that for you!

```
price <- airbnb$price #please use the column name you assigned in problem 3!
service <- airbnb$service.fee #please use the column name you assigned in problem 3!
#now, replace the $ in the vector price and the vector service with an empty string

#write your R code here
price <- str_replace(price, "^\\$", "")
service <- str_replace(service, "^\\$", "")

#attribute the fixed vector back to the dataframe
```

```
#please use the column names you assigned in problem 3!
airbnb$price <- price
airbnb$service.fee <- service
head(price)
```

```
## [1] "966 " "142 " "620 " "368 " "204 " "577 "
```

```
head(service)
```

```
## [1] "193 " "28 " "124 " "74 " "41 " "115 "
```

Problem 5

We want to know the price range for the airbnb (maximum and minimum values), but we want to consider all the costs (price per night plus service fee). The first problem we have is that price and service fees are characters variables, thus, we can't really sum them up. Then, we have to create a new vector with the sum before we take the range information. To find the price range, do the following:

1. Convert the price and service fee columns to numeric values
2. Sum the columns price and service fee, and store the result in a `final_cost` variable
3. Find the minimum and maximum values of `final_cost`. **Note:** because there are NAs in your data, you should ignore them when finding the minimum and maximum values. Check the function parameter called `na.rm`, it should be set to `TRUE` to remove the NA from the results.

```
#write your R code here
```

```
airbnb$price <- as.numeric(airbnb$price)
```

```
## Warning: NAs introduced by coercion
```

```
airbnb$service.fee <- as.numeric(airbnb$service.fee)
```

```
final_cost = airbnb$price + airbnb$service.fee
min(final_cost, na.rm = TRUE)
```

```
## [1] 60
```

```
max(final_cost, na.rm = TRUE)
```

```
## [1] 1199
```