Forums / Programming Assignment 3: Hospital Quality Help Center

Assignment 3, part 4: NAs in output

✉ You are subscribed. Unsubscribe

Assignment3-Part4 × + Add Tag Sort replies by: Oldest first Newest first Most popular

onoharuko · 10 days ago

My rankall function would produce output like (using examples given in the assignment instruction):

```
> head(rankall("heart attack", 20), 10)
                          hospital state
AK                              NA    NA
AL    D W MCMILLAN MEMORIAL HOSPITAL    AL
AR ARKANSAS METHODIST MEDICAL CENTER    AR
AZ               ORO VALLEY HOSPITAL    AZ
CA             SHERMAN OAKS HOSPITAL    CA
CO          SKY RIDGE MEDICAL CENTER    CO
CT           MIDSTATE MEDICAL CENTER    CT
DC                              NA    NA
DE                              NA    NA
FL    SOUTH FLORIDA BAPTIST HOSPITAL    FL
```

or even worse...

```
> tail(rankall("pneumonia", "worst"), 3)
   hospital state
WI       NA    NA
WV       NA    NA
WY       NA    NA
```

I know there are threads in the forum discussing about dealing with NAs in States, but I have tried methods suggested in those threads and still had no idea how to fix my problem. So I thought it would be helpful to share some of my codes (without posting the full code).

1. I first read in the data and created a dataframe containing 5 columns in the original data. I then coerced those outcome columns into numeric and use na.omit to get rid of NAs. I call this dataframe mydataclean:

```
mydataclean <- na.omit(mydata)
> str(mydataclean)
'data.frame':   2709 obs. of  5 variables:
 $ HospitalNames: chr  "SOUTHEAST ALABAMA MEDICAL CENTER" "MARSHALL MEDICAL CENTER SOUTH" "ELI
ZA COFFEE MEMORIAL HOSPITAL" "ST VINCENT'S EAST" ...
```

```
 $ State        : chr  "AL" "AL" "AL" "AL" ...
 $ heart attack : num  14.3 18.5 18.1 17.7 18 15.9 19.6 17.3 17.8 17.5 ...
 $ heart failure: num  11.4 15.2 11.3 10.9 16.6 13.6 12.6 11.8 11.8 10.2 ...
 $ pneumonia    : num  10.9 13.9 13.4 16.2 15.8 10.7 15 9.9 14.3 14.7 ...
 - attr(*, "na.action")=Class 'omit'  Named int [1:1997] 4 5 6 10 13 17 19 23 27 28 ...
  .. ..- attr(*, "names")= chr [1:1997] "4" "5" "6" "10" ...
```

2. Then I checked if the outcome is valid.

3. Then I split the dataframe by state (mysplit), and used lappy to order each groups:

```
listbystate <- lapply(mysplit, function(x) {x[order(x[,outcome],x[,1]),]})
```

4. Then this is what I did with the num argument (I know this if/else looks ugly somehow...):

```
return_rank <- vector('numeric')
        if (num == "best"){return_rank <- 1}
        if (num == "worst"){return_rank <- nrow(mydataclean)}
        else {return_rank <- as.numeric(num)}
```

5. I then used lapply to return the names of the hospitals and the names of states from the ordered list :

```
myorder <- sapply(listbystate, function(x) x[return_rank,1])
myorder2 <- sapply(listbystate,function(x) x[return_rank,2])
```

6. Finally, I cbind the above two lists into finaloutput, changed it into a data frame and added column names. When I tried to test my function using examples provided in the instruction, I got output which I posted at beginning.

**However, When I ran these codes separately (without calling the rankall function) and used "heart attack" as an example, this is what I got for finaloutput:

```
> head(finaloutput)
                              hospital state
AK     PROVIDENCE ALASKA MEDICAL CENTER    AK
AL             CRESTWOOD MEDICAL CENTER    AL
AR              ARKANSAS HEART HOSPITAL    AR
AZ                 MAYO CLINIC HOSPITAL    AZ
CA    GLENDALE ADVENTIST MEDICAL CENTER    CA
CO ST MARYS HOSPITAL AND MEDICAL CENTER    CO
```

So I am assuming the order and rank part is right, and the problem is about returning output?...Can anybody point out for me what is the problem? I appreciated any help!! Thanks.

↑ 0 ↓ · flag

Al Warren · 10 days ago 🔗

"*a dataframe containing 5 columns in the original data. I then coerced those outcome columns into numeric and use na.omit to get rid of NAs.*"

You're omitting NAs before selecting an outcome column. Suppose you're ranking column A. What happens when a row contains a valid number in column A but NA in column B? Select your column first then omit NAs.

1. read the file
2. select three columns - hospital name, state, and one outcome column
3. omit NAs
4. sort by state, outcome, hospital name
5. Process the data.

Also, if you use na.strings="Not Available" in your read function you shouldn't need to coerce.

↑ 4 ↓ · flag

onoharuko · 10 days ago 🔗

Hello Al,

Thanks for your reply. I see what you are saying about why I should not omit NAs before selecting an outcome column. But I do not understand how can I do steps 2,3,4 sequentially.

I select my outcome column when I am ordering the data (using the order function as an argument of the lappy). How am I supposed to first select one outcome column, then omit NAs and then sort?

↑ 1 ↓ · flag

Alireza Abdoli · 10 days ago 🔗

Reading the arguments of order() function (using ?order), you find an argument na.last accepting three values; if na.last = TRUE the NA values in the outcome column are put at the end, if na.last = FALSE then NA values in the outcome column are placed at the top of the data frame, now the interesting part, in case na.last = NA then NA values are removed...

↑ 0 ↓ · flag

ly · 10 days ago 🔗

Why not subset only your outcome column (aside from hospital and state name) as soon as you have validated `outcome`? At that point, you know the only outcome column you need, and you may safely discard other ones from the data.

If you want to do everything within lapply, you could do that too, just by adding more steps to your anonymous function:

```
listbystate <- lapply(mysplit,
    function(x) {
        #omit NAs...
        #then order...
        #then find minimimum...or maximum...etc, if you like.
    })
```

From the looks of it, this timing of your NA removal could be your only issue.

↑ 0 ↓ · flag

Leonard Greski Signature Track · 10 days ago

You also have to set the value of state to the correct state for those states that have fewer hospitals than the required rank number. "best" and "worst" should always return a hospital for 54 states & territories, so there's no need to insert the state codes for states with missing data. A numeric rank of 10, however, should return NA for any states / territories that have 9 or fewer hospitals, so you have to find the states that return <NA> <NA> and reset the state variable to the correct state code (e.g. "DC", "GU", etc.).

regards,

Len

↑ 0 ↓ · flag

onoharuko · 10 days ago

Thank you all for replying...but I guess my problem is still there ;(
@Alireza I tried to use "na.last = NA" but nothing changed
@ly So as you said, I subseted a data frame only containing the hospital names, states and the selected outcome column. I then removed NAs using the na.omit function. Then I sorted by state (using split) and further processed the data(lappy-sapply as I wrote in the post) . (@ Al Warren is this what you are suggesting me to do?) But the output was still as before...

Yesss! I think my only issue is indeed the NAs...but somehow it is just driving me crazy. @ **Leonard Greski** Thanks for your note too! I will look at your comment once I got this removing-NA issue done.

↑ 0 ↓ · flag

onoharuko · 9 days ago

So this is how I subset my only outcome column with hospital names and state names:

```
  if (outcome == "heart attack") {mydata <-data[,c(2,7,11)]}
  if (outcome == "heart failure") {mydata <-data[,c(2,7,17)]}
  if (outcome == "pneumonia") {mydata <-data[,c(2,7,23)]}
```

data is the original data set

I then removed NAs with na.omit

```
mydata <-na.omit(mydata)
```

Is this what I am supposed to do?... I am still getting the same output.

Thank you all again!

↑ 1 ↓ · flag

+ Comment

New post

To ensure a positive and productive discussion, please read our forum posting policies before posting.

B *I* Link <code> Pic Math Edit: Rich ▾ Preview

☐ Make this post anonymous to other students

☑ Subscribe to this thread at the same time

Add post