Lecture 3 problem set

INSERT YOUR NAME HERE

10/11/2019

Extracting and Sorting Data via Tidyverse and base R

The aim of this problem set is to demonstrate there are many different ways to complete the same data management tasks.

Last week you learned to extract variables and observations as well as sort observations the `tidyverse` way via the `select`, `filter`, and `arrange` functions. Lecture 3 demonstrated how some of the tasks done with `tidyverse` functions have a corresponding solution using `base R` syntax.

For the following questions, you'll be asked to complete the same task multiple ways based on the `tidyverse` and `base R` approaches.

Step 1: Remove objects in current R session, load `tidyverse`, and open the data

1. Begin by removing any objects in your current R session by using `rm(list = ls())`. Then load the `tidyverse` library. Lastly, use the `load` function to open the `df_event` dataset via url link

- The url for the `df_event` dataset is https://github.com/ozanj/rclass/raw/master/data/recruiting/recruit_event_somevars.RData
- The data frame `df_event` has one observation for each recruiting event.

Step 2: Extract columns, extract observations, sort observations

Complete all the following questions in three different ways: (1) by using the tidyverse `select`, `filter`, or `arrange` functions, (2) by using base R's subsetting operators, and/or (3) by using base R's `subset` or `order` functions.

I have included rchunks below to indicate how many different ways you should be attempting the tasks.

2. Create a new dataframe by extracting the columns `univ_id`, `event_date`, `event_type`, `zip`, and `med_inc` from `df_event`. Use the names() function to show what columns (variables) are in the newly created dataframe. Print the first 10 observations of the newly created dataframe.

tidyverse

base R using subsetting operators

base R using `subset()`

3. Create a new dataframe from `df_event` that includes recruiting events by the University of Massachusetts Amherst (`univ_id`==166629), that were located at in-state public high schools (`event_type` and `event_state`) where the average median household income (`med_inc`) is equal to or greater than $100,000. Use `nrow` to make sure you are extracting the same number of observations across each approach below.

tidyverse

base R using subsetting operators

base R using `subset()`

4. Create a new dataframe from `df_event` that includes recruiting events by the University of South Carolina Columbia (`univ_id`==218663), that were located at out-of-state public high schools (`event_type` and `event_state`) where the average median household income (`med_inc`) is equal to or greater than $100,000 and the White population in the surrounding area is equal to or greater than 50% of the total population (`pct_white_zip`). Use `nrow` to make sure you are extracting the same number of observations across each approach below.

tidyverse

base R using subsetting operators

base R using `subset()`

5. Create a new dataframe from df_events that sorts by ascending `univ_id`, ascending by `event_date` , ascending `event_state`, descending `pct_white_zip`, descending `med_inc`.

tidyverse

base R using `order()`