Homework 6-Dec

Ben Ridenhour

2022-12-05

Data Manipulation with `dplyr` Practice

This homework is centered around manipulating a data set using functions from the `dplyr` package. You should be able to accomplish everything in the homework using some form of the following functions:

function	description	example
`filter()`	select rows based on some criteria	data %>% filter(age > 20 & sex == "f")
`arrange()`	sort rows	data %>% arrange(age, height)
`select()`	select columns	data %>% select(sex, weight) OR data %>% select(-height)
`rename()`	rename columns	data %>% rename(newName = oldName)
`mutate()`	create new columns	data %>% mutate(old = age > 40)
`case_when()`	recode values	data %>% mutate(ageCat = case_when(age < 20 ~ "young", age < 40 ~ "middle", T~"old"))
`group_by()`	create groups for calculations	data %>% group_by(sex)
`summarize()`	calculate statistics on *groups*	data %>% group_by(sex) %>% summarize(meanW = mean(weight))
`left_join()`	merge data sets on some ID	data %>% left_join(data2, by = "id")

Data for the exercises

We will be using data sets for the following exercises. The first of these comes from the `speff2trial` package. To get the data do the following:

```
install.packages("speff2trial")
library(speff2trial)
trial <- ACTG175
```

If you wish to see a description of the data set use `?ACTG175`. The second set of data are in the Github folder and should be downloaded when you pull the latest version; the file with the data is called `"patient_demo.csv"`. To load the data, make sure your current working directory (use `getwd()` and `setwd()` to check or change, respectively) contains the CSV file then run:

```
demo_data <- read.csv("patient_demo.csv")
```

Once you have loaded both data sets, verify that they exist and look okay within your R global environment.

Exercises

These exercises primarily deal with the `trial` data until later. In each step, be sure to save the data set with the newly manipulated columns.

`rename()`

1. Rename the columns `wtkg` to `weightkg` and `age` to `age_years`.

`mutate()`

2. Create a column called `agem` that displays a patient's age in months rather than years.
3. The column `karnof` shows each patient's Karnofsky score on a scale of 0 to 100. Create a new column called `karnof_b` that has a scale of 0 to 1 instead.
4. Create a column called `sparrow` that has the formula of Karnofsky score/(age + weight).

`arrange()`

5. Sort the `trial` data in ascending values of age, weight, and sparrow, respectively. Verify that the ordering worked.
6. Using the `desc()` function, sort the trial data in order of *descending* weight.

`case_when()`

7. In one *single* call, create two columns in the data: the first creates a column named `gender_char` that shows gender as either "male" or "female" and the second column creates a discrete variable `over50` that is 0 when the patient is younger than 50 and 1 otherwise.

`filter()`

8. Create a new data set `trial_men` that only had data from males.
9. Create another data set `trial_sub` that only has individuals that are: female, haven't used intravenous drugs, and are over 40 years old.

`select()`

10. Drop the `karnof` and `weight` columns from the `trial_men` data and save the results.
11. Keep only the `pidnum`, `treat`, and `cd820` columns from `trial_sub` and save the results.

`left_join()`

12. Merge the `trial` and `demo_data` data sets. Save the resulting data as `trial_full`.
13. Using the `trial_full` data, calculate the mean number of days of exercise reported by patients.

`group_by()` and `summarize()`

14. Group the data by `arms`. Then, for each arm, calculate the mean participant age.
15. Group the data by `arms`. For each arm, calculate the mean participant age and the median Karnofsky score.
16. Group the data by `gender`. Then, separately for male and female patients, calculate the percent who have a history of intravenous drug use. (To calculate a percent of a binary variable with 0s and 1s, just calculate the mean)
17. Group the data by `gender`. Then calculate the percent of male and female patients who have a history of intravenous drug use (drugs), and the mean number of days until a major negative event days
18. Separately for all combinations of `gender` and `race`, calculate the mean age and mean CD4 T cell count at baseline (`cd40`). (Hint: group by gender and race.)

Exercises

19. Now let's check the major differences between the treatment arms. For each arm, calculate the following:

- Mean days until a a major negative event (days)
- Mean CD4 T cell count at baseline.
- Mean CD4 T cell count at 20 weeks.
- Mean CD4 T cell count at 96 weeks.
- Mean change in CD4 T cell count between baseline and 96 weeks
- Number of patients (Hint: use `N = n()`)

20. Repeat the previous analysis, but before you do the grouping and summary statistics, create a new variable called `arms_char` that shows the values of `arms` as characters that reflect what the values actually represent (hint: use `mutate()` and `case_when()`). For example, looking at the help file `?ACTG175`, I can see that the treatment arm of 0 is "zidovudine". I might call this arm `"Z"`. Do this **all in the same chunk of code**.

21. Repeat the previous analysis, but only include patients with a Karnofsky score equal to 100, and who did not use zidovudine in the 30 days prior to the treatment initiation (`z30`).

22. In **one block of code**, complete the following tasks

- Change the name of the column `weightkg` to `weight`
- Create a new column called `drugs_char` that uses strings instead of numbers to indicate drug use
- Filter the data to only include male patients with id numbers greater than 10100
- Group the data by `drugs_char` and `arms`
- For each group calculate the mean age, percentage of patients that are male, and mean number of days until a major negative event