Skills Problem Set IV

Fernanda Sobrino

4/5/2021

Due Thursday May 6, midnight Central Time.

Submit to add_link here

Name your submission files `skills_ps_4.Rmd` and `skills_ps_4.pdf` **(5 points)**.

Your code should adhere to the style guide. (`styler` is your friend.) **(10 points)**.

Remember to map your answers with gradescope! **(We will take 5 points if you do not do this)**

This submission is my work alone and complies with the 30535 integrity policy.

Add your initials to indicate your agreement: **___**

Add names of anyone you discussed this problem set with: **___**

Late coins used this pset: 0. Late coins left after submission: X.

1 Tidy

1.1 Tidy data with `pivot_wider()` and `pivot_longer()` (25 points)

1. The data set `billboard` inside `tidyr`. Has the song rankings for Billboard top 100 in the year 2000. Is this data tidy? If not, identify the problem and solve it. Be careful with missing values, do we need them in the final data set or not?

2. The data set `fish_encounters` inside `tidyr` shows information about different monitors that capture fish swimming down a river.

    1. Is this data tidy? If not, identify the problem and solve it.
    2. Which kind of missing values does this data has? What do they mean?

3. The data set `us_rent_income` inside `tidyr` shows income and rent by estate in 2017 from the American Community Survey. Is this data tidy? If not, identify the problem and solve it. How is this case different from the ones we have seen so far?

4. `pivot_longer()` and `pivot_wider()` are not perfectly symmetrical. Carefully consider the following example. Why do we need quotes on the arguments `names_to` and `values_to`, but not in `names_from` and `values_from`?

```
soccer <- tibble(
  game = c("Real Sociedad", "Real Sociedad", "Huesca", "Huesca"),
  player = c("Messi", "Griezmann", "Messi", "Griezmann"),
  goals = c(2,1,2,1)
)
soccer %>%
  pivot_wider(names_from = player, values_from = goals) %>%
  pivot_longer(Messi:Griezmann,
```

```
                  names_to = "player",
                  values_to = "goals")
```

5. This code fails. Explain the error message. How could it be fixed?

```
table4a %>%
  pivot_longer(1999:2000,
               names_to = "year",
               values_to = "cases")
```

6. Why does `pivot_wider` fails on this tibble? And a new column to address the problem and show that `pivot_wider` works on your new update dataset.

```
soccer <- tribble(
  ~player,    ~game, ~goals,
  "Messi",    "Real Sociedad", 2,
  "Messi",     "Huesca",  2,
  "Messi",     "Real Sociedad", 0,
  "Messi",     "Huesca", 1,
  "Griezmann", "Real Sociedad", 1,
  "Griezmann", "Huesca", 1
)
```

7. Tidy the pivot table below. Do you need to make it wider or longer? What are the variables?

```
preg <- tribble(
  ~pregnant, ~male, ~female,
  "yes", NA, 10,
  "no", 20, 12
)
```

8. What do the `extra` and `fill` arguments do in `separate()`? Hint: experiment with the various options for the following two data sets

```
tibble(x = c("a,b,c", "d,e,f,g","h,i,j")) %>%
  separate(x, c("one", "two", "three"))
tibble(x = c("a,b,c","d,e","f,g,i")) %>%
  separate(x, c("one","two","three"))
```

1.2 tidying case study (30 pts)

1. In this WHO case study in Ch 12.6 Hadley set `na.rm = TRUE` just to make it easier to check that we had the correct values.
    1. Are there implicit missing values? Use a command you learned in the tidy data slides/videos. If there are implicit missing values, how many rows? If not, show how you know that there are not.
    2. How many country-year pairs are explicitly missing TB data?
2. In this WHO case study in Ch 12.6, what's the difference between an NA and zero?
3. What happens if you neglect the `mutate()` step?
4. Health outcomes are often sexed. As in certain maladies are more associated with males or females. Using the tidied WHO data, you will make an informative visualization to address the question:"To what extent is Tuberculosis associated with a specific sex and has this changed from 1997 onward?" (follow the steps closely and answer where there is a question.)
    1. For each country, year, and sex compute the total number of cases of TB.
    2. Using raw values is probably not going to provide clear evidence. Why not?

3. For each country-year, compute the ratio of male to female patients.
4. Producing these ratios by year (ignoring county) is probably a bad idea. Why?

Result: 1. Make a plot that address the main question (To what extend is tuberculosis associated with a specific sex and has this changed from 1997 onward?). Think carefully which kind of plot you are going to use, you want to uncover the general pattern but also learn specifics about your data. 1. Write a quick summary of lessons learned from your final data visualization. What is the general conclusion from this plot? Did you find any other valuable information from your plot?

1.3 Unseen untidy data (15 pts)

1. The data set `world_bank_pop` is messy. Tidy it, show each of your steps and at the end write a short paragraph of what you just did. Your final data should look like this:

```
## # A tibble: 4,752 x 6
##    country year  URB_TOTL URB_GROW POP_TOTL POP_GROW
##    <chr>   <chr>    <dbl>    <dbl>    <dbl>    <dbl>
##  1 ABW     2000     42444   1.18      90853   2.06
##  2 ABW     2001     43048   1.41      92898   2.23
##  3 ABW     2002     43670   1.43      94992   2.23
##  4 ABW     2003     44246   1.31      97017   2.11
##  5 ABW     2004     44669   0.951     98737   1.76
##  6 ABW     2005     44889   0.491    100031   1.30
##  7 ABW     2006     44881  -0.0178   100832   0.798
##  8 ABW     2007     44686  -0.435    101220   0.384
##  9 ABW     2008     44375  -0.698    101353   0.131
## 10 ABW     2009     44052  -0.731    101453   0.0986
## # ... with 4,742 more rows
```

2 Data Types Strings (15 pts)

Hint: try lots of test cases to be sure you get it right

1. Write a regular expression to match any superhero name that ends with `man` and that is shorter or equal to 8 characters. This means it should be able to match `Batman` but not `Spiderman` (be careful I don't want it to match a regular `man`). Prove your regular expression works with three examples.
2. Given the corpus of fruits in `stringr::fruit`, create regular expressions that find all fruits that:
    1. Ends with "t".
    2. Starts with "h"
    3. Are exactly 6 letters long. (Don't use `str_length()`!)
    4. Are 10 letters or longer. (Note: including all the output here would make grading difficult. instead, use `sum(str_detect(stringr::words, regex))` to count the number of strings that match each of the patterns above)
3. Create regular expressions to find all words in `stringr::words` that meet the following criteria. In addition, please provide two test cases where your regular expression returns a match and two test cases that do not return a match.
    1. Start with an `a` or a `o`.
    2. That only contain consonants. (Hint: thinking about matching "not"-vowels.)
    3. That are berries, i.e contain the word `berry`
    4. End with `ine` or `een`.
4. Show how telephone numbers are written in your country with three examples. Create a regular expression that will match telephone numbers as commonly written in your country.