Final Assessment due Jun 15, 2021 14:59 +03

Question 1

1/1 point (graded)

In the `extdata` directory of the **dslabs** package, you will find a PDF file containing daily mortality data for Puerto Rico from Jan 1, 2015 to May 31, 2018. You can find the file like this:

```
fn <- system.file("extdata", "RD-Mortality-Report_2015-18-180531.pdf", pac
```

Find and open the file or open it directly from RStudio. On a Mac, you can type:

```
      system2("open", args = fn)
```

and on Windows, you can type:

```
      system("cmd.exe", input = paste("start", fn))
```

Which of the following best describes this file?

○ It is a table. Extracting the data will be easy.

○ It is a report written in prose. Extracting the data will be impossible.

● It is a report combining graphs and tables. Extracting the data seems possible.

○ It shows graphs of the data. Extracting the data will be difficult.

✓

Submit You have used 1 of 2 attempts

ℹ Answers are displayed within the problem

Question 2

1/1 point (graded)

We are going to create a tidy dataset with each row representing one observation. The variables in this dataset will be year, month, day and deaths.

Use the **pdftools** package to read in `fn` using the `pdf_text()` function. Store the results in an object called `txt`.

Describe what you see in `txt`.

○ A table with the mortality data.

● A character string of length 12. Each entry represents the text in each page. The mortality data is in there somewhere.

○ A character string with one entry containing all the information in the PDF file.

○ An html document.

✓

Answer code

```
        txt <- pdf_text(fn)
```

Submit You have used 1 of 2 attempts

ℹ Answers are displayed within the problem

Question 3

3/3 points (graded)

Extract the ninth page of the PDF file from the object `txt`, then use the `str_split()` function from the **stringr** package so that you have each line in a different entry. The new line character is `\n`. Call this string vector `x`.

Look at `x`. What best describes what you see?

○ It is an empty string.

○ I can see the figure shown in page 1.

○ It is a tidy table.

● I can see the table! But there is a bunch of other stuff we need to get rid of.

✓

Answer code

```
            x <- str_split(txt[9], "\n")
```

What kind of object is `x`?

list ✓ **Answer:** list

Answer code

```
            class(x)
```

How many entries does `x` have?

1 ✓ **Answer:** 1

1

Answer code

```
          length(x)
```

Submit

You have used 1 of 5 attempts

ℹ Answers are displayed within the problem

Question 4

2/2 points (graded)

Define `s` to be the first entry of the `x` object.

What kind of object is `s`?

character vector ⌄ ✓ **Answer:** character vector

Answer code

```
          s <- x[[1]]
class(s)
```

How many entries does `s` have?

40 ✓ **Answer:** 40

40

Answer code

```
          length(s)
```

Submit

You have used 1 of 5 attempts

ℹ Answers are displayed within the problem

Question 5

1/1 point (graded)
When inspecting the string we obtained above, we see a common problem: white space before and after the other characters. Trimming is a common first step in string processing. These extra spaces will eventually make splitting the strings hard so we start by removing them.

We learned about the command `str_trim()` that removes spaces at the start or end of the strings. Use this function to trim `s` and assign the result to `s` again.

After trimming, what single character is the last character of element 1 of `s`?
Your answer should be one character.

s

✓ **Answer:** s

Answer code

```
          s <- str_trim(s)
s[1]    # print string, visually inspect last character
```

Submit

You have used 1 of 10 attempts

ℹ Answers are displayed within the problem

Question 6

1/1 point (graded)
We want to extract the numbers from the strings stored in `s`. However, there are a lot of non-numeric characters that will get in the way. We can remove these, but before doing this we want to preserve the string with the column header, which includes the month abbreviation.

Use the `str_which()` function to find the row with the header. Save this result to `header_index`. Hint: find the first string that matches the pattern `"2015"` using the `str_which()` function.

What is the value of `header_index`?

2

✓ **Answer:** 2

2

Answer code

```
         header_index <- str_which(s, "2015")[1]
header_index
```

Submit

You have used 1 of 10 attempts

ℹ Answers are displayed within the problem

Question 7

2/2 points (graded)

We want to extract two objects from the header row: `month` will store the month and `header` will store the column names.

Save the content of the header row into an object called `header`, then use `str_split()` to help define the two objects we need.

What is the value of `month`?

Use `header_index` to extract the row. The separator here is one or more spaces. Also, consider using the `simplify` argument.

SEP

✓ **Answer:** SEP

Answer code

```
            tmp <- str_split(s[header_index], "\\s+", simplify = TRUE)
month <- tmp[1]
header <- tmp[-1]
month
```

What is the third value in `header`?

2017

✓ **Answer:** 2017

2017

Answer code

```
            header[3]
```

Submit

You have used 1 of 10 attempts

ℹ Answers are displayed within the problem

This assessment continues on the next page. Make sure the variable `s` is defined as in the exercises above.