**Assessment Part 2: String Processing Part 3**

Import raw Brexit referendum polling data from Wikipedia:

library(rvest)

library(tidyverse)

library(stringr)

url <- "https://en.wikipedia.org/w/index.php?title=Opinion\_polling\_for\_the\_United\_Kingdom\_European\_Union\_membership\_referendum&oldid=896735054"

tab <- read\_html(url) %>% html\_nodes("table")

polls <- tab[[5]] %>% html\_table(fill = TRUE)

You will use a variety of string processing techniques learned in this section to reformat these data.

**Question 5**

1/1 point (graded)

Some rows in this table do not contain polls. You can identify these by the lack of the percent sign (%) in the Remain column.

Update polls by changing the column names to c("dates", "remain", "leave", "undecided", "lead", "samplesize", "pollster", "poll\_type", "notes") and only keeping rows that have a percent sign (%) in the remain column.

How many rows remain in the polls data frame? correct

129 Loading

You have used 3 of 10 attempts Some problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

**Question 6**

3.0/3.0 points (graded)

The remain and leave columns are both given in the format "48.1%": percentages out of 100% with a percent symbol.

Which of these commands converts the remain vector to a proportion between 0 and 1?

Check all correct answers.

as.numeric(str\_remove(polls$remain, "%"))

as.numeric(polls$remain)/100

parse\_number(polls$remain)

str\_remove(polls$remain, "%")/100

as.numeric(str\_replace(polls$remain, "%", ""))/100

parse\_number(polls$remain)/100

correct

You have used 1 of 3 attempts Some problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

**Question 7**

3/3 points (graded)

The undecided column has some "N/A" values. These "N/A"s are only present when the remain and leave columns total 100%, so they should actually be zeros.

Use a function from **stringr** to convert "N/A" in the undecided column to 0. The format of your command should be function\_name(polls$undecided, "arg1", "arg2").

What function replaces function\_name? correct

What argument replaces arg1?

Omit the quotation marks.

correct

What argument replaces arg2?

Omit the quotation marks.

correct

You have used 1 of 10 attempts Some problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

**Question 8**

1.9949999999999999/3.5 points (graded)

The dates column contains the range of dates over which the poll was conducted. The format is "8-10 Jan" where the poll had a start date of 2016-01-08 and end date of 2016-01-10. Some polls go across month boundaries (16 May-12 June).

The end date of the poll will always be one or two digits, followed by a space, followed by the month as one or more letters (either capital or lowercase). In these data, all month abbreviations or names have 3, 4 or 5 letters.

Write a regular expression to extract the end day and month from dates. Insert it into the skeleton code below:

temp <- str\_extract\_all(polls$dates, \_\_\_\_\_)

end\_date <- sapply(temp, function(x) x[length(x)]) # take last element (handles polls that cross month boundaries)

Which of the following regular expressions correctly extracts the end day and month when inserted into the blank in the code above?

Check all correct answers.

"\\d?\\s[a-zA-Z]?"

"\\d+\\s[a-zA-Z]+"

"\\d+\\s[A-Z]+"

"[0-9]+\\s[a-zA-Z]+"

"\\d{1,2}\\s[a-zA-Z]+"

"\\d{1,2}[a-zA-Z]+"

"\\d+\\s[a-zA-Z]{3,5}"

partially correct

Answer

Incorrect:

Try again. This only extracts the last digit of the day and the first letter of the month.

Try again. This only extracts the first letter of the month.

Try again. This omits the space between the day and month.

You have used 3 of 3 attempts Some problems have options such as save, reset, hints, or show answer. These options follow the Submit button.