Audit Access Expires Mar 19, 2020
You lose all access to this course, including your progress, on Mar 19, 2020.

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_r
eferendum&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

0/1 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?

0

✗ **Answer:** 129

\(\)

Answer code

```
names(polls) <- c("dates", "remain", "leave", "undecided", "lead", "samplesize", "pollster", "poll_type"
polls <- polls[str_detect(polls$remain, "%"), -9]
nrow(polls)
```

Submit You have used 10 of 10 attempts

Answers are displayed within the problem

Question 6

0.51/3 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` ✓

✱

Answer

Incorrect:

Try again. These values are between 0 and 100, but you want a proportion between 0 and 1.

Try again. You need to remove the "%" symbol.

Try again. Make sure the final result is a numeric type.
Try again. At least one of the answers you didn't choose is correct.
Try again. At least one of the answers you didn't choose is correct.

Submit You have used 3 of 3 attempts

ℹ Answers are displayed within the problem

Question 7

1/3 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`?

0

✗ **Answer:** str_replace **or** str_replace_all **or** str_replace() **or** str_replace_all()

What argument replaces `arg1`?
Omit the quotation marks.

0

✗ **Answer:** N/A

What argument replaces `arg2`?
Omit the quotation marks.

0

✓ **Answer:** 0

Submit You have used 10 of 10 attempts

ℹ Answers are displayed within the problem

Question 8

1.015/3.5 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 b
```

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}"` ✓

*

Answer

Incorrect:

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

Try again. At least one of the answers you didn't choose is correct.

Try again. At least one of the answers you didn't choose is correct.

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

Try again. At least one of the answers you didn't choose is correct.

Submit You have used 3 of 3 attempts

ℹ Answers are displayed within the problem