**Assessment: Web Scraping**

**Introduction: Questions 1-3**

Load the following web page, which contains information about Major League Baseball payrolls, into R: [https://web.archive.org/web/20181024132313/http://www.stevetheump.com/Payrolls.htm](https://web.archive.org/web/20181024132313/http:/www.stevetheump.com/Payrolls.htm)

library(rvest)

url <- "https://web.archive.org/web/20181024132313/http://www.stevetheump.com/Payrolls.htm"

h <- read\_html(url)

We learned that tables in html are associated with the table node.  Use the html\_nodes function and the table node type to extract the first table. Store it in an object nodes:

nodes <- html\_nodes(h, "table")

The html\_nodes function returns a list of objects of class xml\_node. We can see the content of each one using, for example, the html\_text function. You can see the content for an arbitrarily picked component like this:

html\_text(nodes[[8]])

If the content of this object is an html table, we can use the html\_table function to convert it to a data frame:

html\_table(nodes[[8]])

You will analyze the tables from this HTML page over questions 1-3.

**Question 1**

2.0/2.0 points (graded)

Many tables on this page are team payroll tables, with columns for rank, team, and one or more money values. Note that not all tables have the same column names.

Convert the first three tables in nodes to data frames and inspect them.

Which of the first three nodes are tables of team payroll?

Check all correct answers. Look at table content, not column names.

None of the below

Table 1

Table 2

Table 3

correct

You have used 1 of 2 attempts Some problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

**Question 2**

2.0/2.0 points (graded)

For the last 3 components of nodes, which of the following are true? (Check all correct answers.)

Check all correct answers.

All three entries are tables.

All three entries are tables of payroll per team.

The last entry shows the average across all teams through time, not payroll per team.

None of the three entries are tables of payroll per team.

correct

You have used 1 of 2 attempts Some problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

**Question 3**

1/1 point (graded)

Create a table called tab\_1 using entry 10 of nodes. Create a table called tab\_2 using entry 19 of nodes.

Note that the column names should be c("Team", "Payroll", "Average"). You can see that these column names are actually in the first data row of each table, and that tab\_1 has an extra first column No. that should be removed so that the column names for both tables match.

Remove the extra column in tab\_1, remove the first row of each dataset, and change the column names for each table to c("Team", "Payroll", "Average"). Use a full\_join by the Team to combine these two tables.

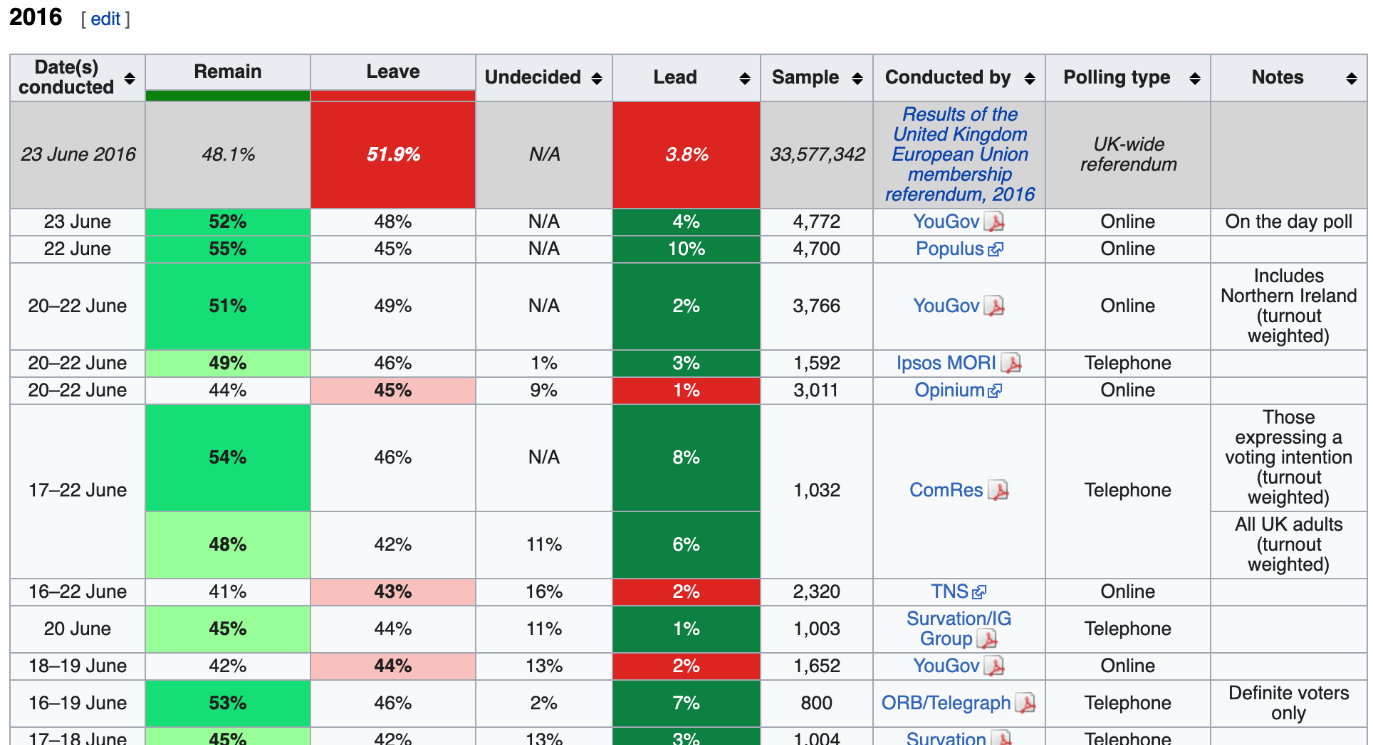
How many rows are in the joined data table? correct

58 Loading

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.

**Introduction: Questions 4 and 5**

The Wikipedia page on [opinion polling for the Brexit referendum](https://en.wikipedia.org/w/index.php?title=Opinion_polling_for_the_United_Kingdom_European_Union_membership_referendum&oldid=896735054), in which the United Kingdom voted to leave the European Union in June 2016, contains several tables. One table contains the results of all polls regarding the referendum over 2016:



Use the **rvest** library to read the HTML from this Wikipedia page (make sure to copy both lines of the URL):

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

**Question 4**

1/1 point (graded)

Assign tab to be the html nodes of the "table" class.

How many tables are in this Wikipedia page? correct

38 Loading

You have used 2 of 10 attempts Some problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

**Question 5**

1/1 point (graded)

Inspect the first several html tables using html\_table with the argument fill=TRUE (you can read about this argument in the documentation). Find the first table that has 9 columns with the first column named "Date(s) conducted".

What is the first table number to have 9 columns where the first column is named "Date(s) conducted"? correct

5 Loading

You have used 2 of 10 attempts