



7316 - INTRODUCTION TO R

Module 3: Exercise

Teacher: Mickaël Buffart (mickael.buffart@hhs.se)

TABLE OF CONTENTS

1	Scraping data from a website	1
2	String manipulation	1
3	More string manipulations	2
4	Writing a function that can be re-used.....	2
5	Automating this process	2

In this assignment, you will create your own dataset by scraping a website, manipulating strings, merging the dataset with another dataset and finally write your own routine to analyze the data.

1 SCRAPING DATA FROM A WEBSITE

Go on the website <https://www.beeradvocate.com/beer/top-rated/>. BeerAdvocate is a website that crowd-sources beer reviews and ratings.

1. Save the URL as a string.
2. using `read_html()` and `html_table()` from the `rvest` package to extract the table with the reviews (use the header option)
3. view the structure of the table.
4. Clean up the table:
 - Extract the data.frame from the list
 - Drop variables that contain no information ("Yours")
 - Change the class of the variables to their appropriate format. *Hint:* You will need to remove the commas from the numbers > 1000 in the Ratings variable to get them into a format that can be converted into a numerical. You can use `str_replace()` or `gsub()`
 - Rename the variables with appropriate names.

You should in the end have a table with 4 variables. We will deal with the information in the string in the next section.

2 STRING MANIPULATION

We now have a table with 4 variables but the "Name" variable contains information that we would like to have separately. But first we must take care of missing values. Some beers

do not have information about the alcohol content. We can identify them by looking for the percentage sign in the string variable.

5. Find all observations that contain a % sign. Filter the data according to this criterion.
6. Split the alcohol content from the string, trim it and store it into a separate, numeric variable. You can use `strsplit()` or `gsub()` to split the string along "|". Use a double escape symbol "\\|", otherwise R cannot interpret the character you want to split by.
7. Unlist the resulting list.
8. save the length of the unlisted vector as a variable `n`.
9. create a $n/2 \times 2$ matrix and fill it with the unlisted data.

Now you should have a table with two variables that can be further manipulated before we put it back into the original dataset.

3 MORE STRING MANIPULATIONS

10. Rename the columns "Name" and "Percentage"
11. Trim the space from the alcohol percentage
12. Remove the percentage sign using `str_replace()` or `gsub()` and a double escape (as with the "|" symbol)
13. convert the Percentage value into a numerical
14. Replace the old Name variable with the new one and add the Percentage column to the original table.

At this point you should have a nice dataset with 5 variables.

4 WRITING A FUNCTION THAT CAN BE RE-USED.

Since we want to automate this process, it would be convenient to have a function that performs the step of trimming and turning the percentage into a numerical.

15. Write a function that takes a character vector as an input.
16. Inside the function, trim the left side, remove the percentage sign and turn it into a numeric

5 AUTOMATING THIS PROCESS

If you select different countries for the BeerAdvocate ratings, you will notice that the URL stays the same but a country code is added.

We now want to extract the ratings for Australia, Austria, Germany, Sweden, the UK and the United States.

17. create a vector with the country codes "de", "se" and "uk".
18. Write a loop that loops through this vector.
19. In each iteration, the loop should add the country code to the url and extract the html table, as you did above.
20. Rename the columns as "Rank", "Name", "Ratings", "Avg", "Yours"

21. If Ratings is read as an integer, convert it into a numeric (remove the comma, as above). *Hint: use `is_character()` in an if-statement.*
22. In addition, you should add the country code as a variable “Country” to the table.
23. If you are in the first iteration, save this html table as a data.frame, call it `data_combined` For each following iteration, simply append the new data.
24. Remove the “Yours” variable
25. Filter out all observations where the string variable does not contain a percentage sign.
26. Perform the same procedure as above to get the alcohol percentage out of the Name string in `data_combined`