Contents

Description

Exercise 1: Preparation (3 points)

Exercise 2: Fixing the header (2 points)

Exercise 3: Removing additional information (2 points)

Exercise 4: Gathering variables (2 points)

Exercise 5: Fixing class types (2 points)

Exercise 6: Sorting the data (2 points)

Exercise 7: Future improvements (2 points)

The datasets are in comma-separated value (CSV) and originate from the study of Peleg et al. (2008; <https://doi.org/10.1007/s11104-007-9417-z>)

We think that the table is a bit “messy” and the goal of this coursework is to clean the data throughout a series of exercises.

Please make use of the library tidyverse only.

Exercise 1: Preparation (3 points)

Find your table in the paper (hint: look at the name of the file).

• 1a. What does the table show? Brief answer, you don’t need to read the whole paper.

• 1b. Did we do any modification (i.e. cut some columns, changed names, etc. . . ) to the original table? Which (if the case)?

• 1c. Do you consider that the table is “messy”? Why?

Exercise 2: Fixing the header (2 points)

The table contains a “double header”. Make the corresponding modifications to merge them into one single header.

Exercise 3: Removing additional information (2 points)

Some rows do not display values corresponding to the column but some additional information (e.g., \*\*\*, Wild emmer wheat, Durum wheat, . . . ), remove those rows.

Tip: reading the data with the option na.strings="" might help you here

Exercise 4: Gathering variables (2 points)

Gather the data in order to have one column with the name of the different variables that were written as “second header” and another column that contains the value of those variables.

Look at the example provided at the end of the document.

Exercise 5: Fixing class types (2 points)

Check the class type of each column (character, integer, numeric, etc..)

• 5a. Do you see any class that does fit well to the type of values? Which one(s)? (Hint: use str())

• 5b. Change the class type of these columns to fit them to type of values that they have.

Exercise 6: Sorting the data (2 points)

Sort the data according to the first column of your table.

Exercise 7: Future improvements (2 points)

Now the data looks better, right? Unfortunately during the process we have dropped some information. . .

• 7a. Which information have we lost in the way of cleaning the data?

• 7b. Make a suggestion of how would you further improve or clean the data and explain why. Do not write the code, just explain what else would you show in the table and/or how would you further clean the data and explain why.

Assignment should be submitted in RMarkdown format. I need both Rmd and .pdf (or .html) files.

We will re-run your code, so make sure it is correct and does not have errors

You can clean the data in one single step or in different steps, both ways are equally fine.

For the writing questions: please write short and concise answers.

* This work will be checked for plagiarism.

