**Week 45**

1. **Answer the question of: "What are basic principles for using spreadsheets for good data organisation?" (no more than 250 words)**

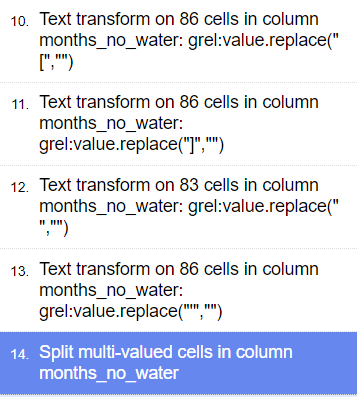
* Put the variables in columns
* Put all of the different observations in different rows
* Refrain from putting different pieces of data in one cell
* Don’t mess with the raw data
* After you are done, export the data to a format, that is accessible for many different people.

1. **Does OpenRefine alter the raw data during sorting and filtering?**

* I think the answer is no, because you are able to go back and undo your steps. Thus OpenRefine doesn’t alter the raw data.

1. **Fix the**[**interviews dataset**](https://ndownloader.figshare.com/files/11502815)**in OpenRefine enough to answer this question: "Which two months are reported as the most water-deprived/driest by the interviewed farmer households?"**

* First I looked at all the the columns, and I found “months\_no\_water” to be the one I should use.
* I use the following regular expressions to make the data in the column readable to the program.



And then I make a text facet, and sort by count. I get the following result:



October and September are the most water deprived months

1. **Finish the tutorial and report what village \*you\* think hides behind the number '49'.**

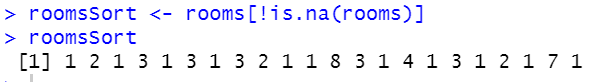
* I have followed the tutorial, and based on GPS\_Lattitude and GPS\_Longitude ‘49’ had very similar coordinates to Chirodzo, therefore I think it is Chirodzo.
* Besides, the person from ‘49’, is interviewed on the same date as people from Chirodzo.

**Week 47**

1) Use R to figure out how many elements in the vector below are greater than 2 . (You need to filter out the NAs first)

rooms <- c(1, 2, 1, 3, 1, NA, 3, 1, 3, 2, 1, NA, 1, 8, 3, 1, 4, NA, 1, 3, 1, 2, 1, 7, 1, NA)

* First I filter out the NA’s, and make a new vector called roomsSort



* I then sort it by using the following code:



* There are 8 elements in the vector that are bigger than 2.

2) What is the average number of rooms (result of running mean() function) in the above 'rooms' vector? Again, best remove the NAs first.

* I run the mean on my vector, where I have already sorted out the NA’s



* The average element in rooms is 2,31.

3) What type of data is in the 'rooms' vector? What function helps you determine the answer?

* I use the following vector:



* It is numeric, because R recognices that ‘NA’ isn’t a character in this instance.

4) Inside your R Project (.Rproj), install the 'tidyverse' package and use the download.file() and read\_csv() function to read the SAFI\_clean.csv dataset into your R project as 'interviews' digital object (see instructions in https://datacarpentry.org/r-socialsci/setup.html and 'Starting with Data' section). **Take a screenshot**of your RStudio interface showing a) the script you used to create the object, b) the 'interviews' object in the Environment and the c) structure of your R project in the bottom right Files pane. Save the screenshot as an image and put it in your**AUID\_lastname\_firstname** repository inside our Github organisation ([github.com/Digital-Methods-HASS](https://www.github.com/Digital-Methods-HASS)). In this Blackboard submission, list the URL leading to the screenshot in your github repository.

* <https://github.com/Digital-Methods-HASS/au613487_Jakob_Trans/blob/main/screenshot.png>

**Week 44**

1: What regular expressions do you use to extract all the dates in this blurb: http://bit.ly/regexexercise2 and to put them into the following format YYYY-MM-DD ?

* Regerxercise2:
  + I copy-paste the text into regex101.com
  + I put in the following code: (\d+).(\d+)..?(\d{4}). This code matches all the different formats of dates, and puts them into 3 groups.
  + I want to match the format YYYY-MM-DD. To do this, I Have to flip the format from the ‘Test String’, which is generally something along the lines of MM-DD-YYYY. I therefore put the following in ‘Substitution’: $3-$1-$2. I now have the correct setup.

2: Write a regular expression to convert the stopwordlist from Voyant in http://bit.ly/regexexercise3 into a neat stopword list for R (where words are surrounded by "word" and separated by commas, such as http://bit.ly/regexexercise4). Then take the stopwordlist from R http://bit.ly/regexexercise4 and convert it into a Voyant list (words on separate line without interpunction)

* Regexercise3:
  + First i copied all the text from excel into regex101.com
  + Then i used “\n”, which deletes all new lines. Now i just need to put quotations, and commas between words. I enter the following into ‘substitution’: “, “
  + I’m missing quotations at the very start and at the end. I put them in manually. Now i have converted Voyant to R.
* Regexercise4:
  + I Copy paste the text into regex101.com
  + I put in the regular expression ‘\”, ‘
  + I put in ‘\n’ in substitution, to make a new line.
  + I then copy paste the substituted text into the ‘test string’, and put in \”
  + I have now converted from R to Voyant.