Learning journal II – Data organization in spreadsheets

Marie Mortensen

9/21/2020

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

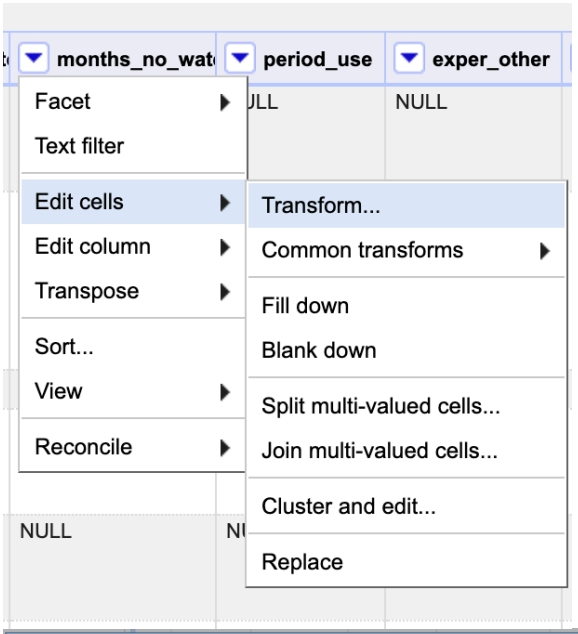
The ways to achieve good dataorginazation in spreadsheets can be challenging when one has already acquired habits and self-made principles of organizing data. However, following some general principles can make spreadsheets and code a lot easier to handle e.g. when you return to a spreadsheet half a year after you made it and forgot your self-made principles. The basic principles involve having one row per observation and being consistent about observation categories such as dates but as well, say, gender and instead of writing "male", "woman", "female", choose one description of each category (Bromana, Woo, 2017). Another principle is avoid leaving cells empty; cells might not have any observation but here you can use NA to clarify that this is the case. Additionally, a principle involves creating a data dictionary which refers to making "meta-data" such as information about variables and description of measurement units. Avoid changing the font of the text or highligt observations as it can both be misleading and uninterpretable and might get lost. In relation to saving variables and data some good principles involve making meaningful names for columns, make sure to make backups and do not overwrite old data, use data validation (making e.g. excel aware that there may only be e.g. positive values in a column) and save data as .txt.

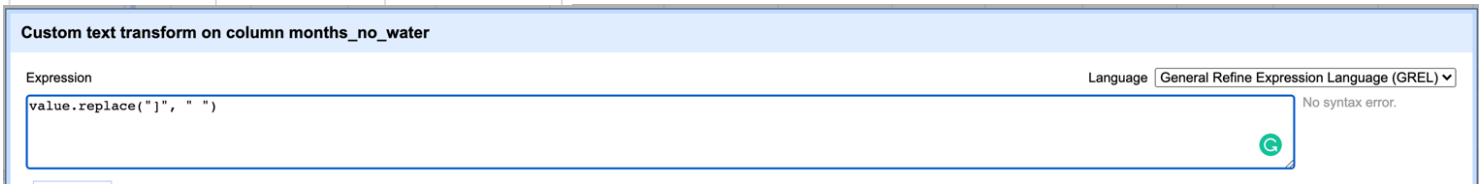
2 - Does OpenRefine alter the raw data during sorting and filtering? In OpenRefine it is possible to sort and fiter and perform various changes without affecting the raw data. However one can save the cleaned/sorted/filtered data as a new file.

3 - Fix the interviews dataset in OpenRefine enough to answer this question: "Which two months are reported as the most water-deprived/driest by the interviewed farmer households?"

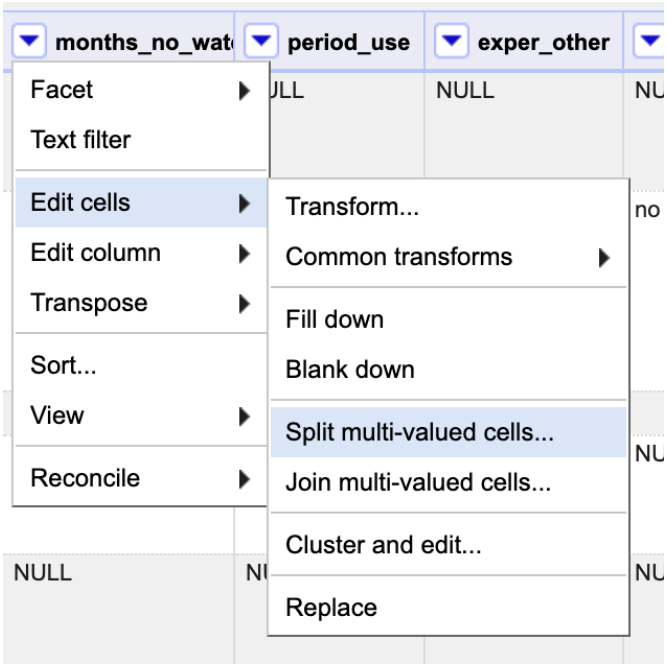
Below, I will describe the steps from selecting a dataset to work with to having summarized a data set such that it answers the question above.

Firstly, In the section “Create project” I choose the csv file “SAFI_openrefine”. I see that the column “months_no_water” contains values where several months are included in one row seperated by ; and enclosed in “[]”, and therefore I’am thinking to change this by the value.replace (which i could not remember the name of and looked up on Google). I try to remove the hard brackets in order to summarize the data.





Afterwards, I try to use the function where rows are split up into more choosing the “split multivalues column” pane and defining ; as the seperator.



As a result, values are split into multiple rows but I can see that some months appear in more than one category and I'm thinking this could be because there are spaces before some of them. I could perhaps have removed the spaces before I split the rows up. Or else I can try to use cluster function. I choose to go back and remove spaces before i split the values.

I'm undoing my step by going to the "undo/redo" section and pressing the section I want to return to.

Now I'm making an additional transformation with value replace writing value.replace(" ","") and thereby removing whitespaces.I try to split columns again. Now the pane with months without water has only one category per month which is how I wanted it ☺ Lastly, to find the months of most water deprivation I choose to sort the column by"count" instead of "name" and this shows me that October and September were the driest months with 74 and 70 mentions.


Facet / Filter
Undo / Redo 4 / 4
Refresh
Reset All
Remove All
months_no_water
change
11 choices Sort by: name count
Cluster
'Oct' 74
'Sept' 70
'Nov' 51
NULL 45
'Aug' 33
'Dec' 11
'Jan' 2
'July' 2
'Apr' 1
'June' 1
'May' 1
Facet by choice counts