W2: Open Refine

1. Answer the question of: "What are basic principles for using spreadsheets for good data organization?"
   a. Don't trust Excel ☺
   b. Do not put different tables/data frames in the same document – you would rather have one dataset you can then convert to R
   c. Your table is supposed to be readable for both human and computers, so don't use highlighting or visual changes when you write (your computer can't read a red color)
   d. Don't use dot or commas or special characters (æ,ø,å e.g.).
   e. Use underscore _ instead of space
   f. Your fields should not be empty if you lack data. Write NA (No answer). If a field is empty you don't know if it has just not been filled out

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

No, it does not. OpenRefine does not do anything with the dataset, rather it filters it and shows different facets of it. You can actively change the data in the "transform" function.

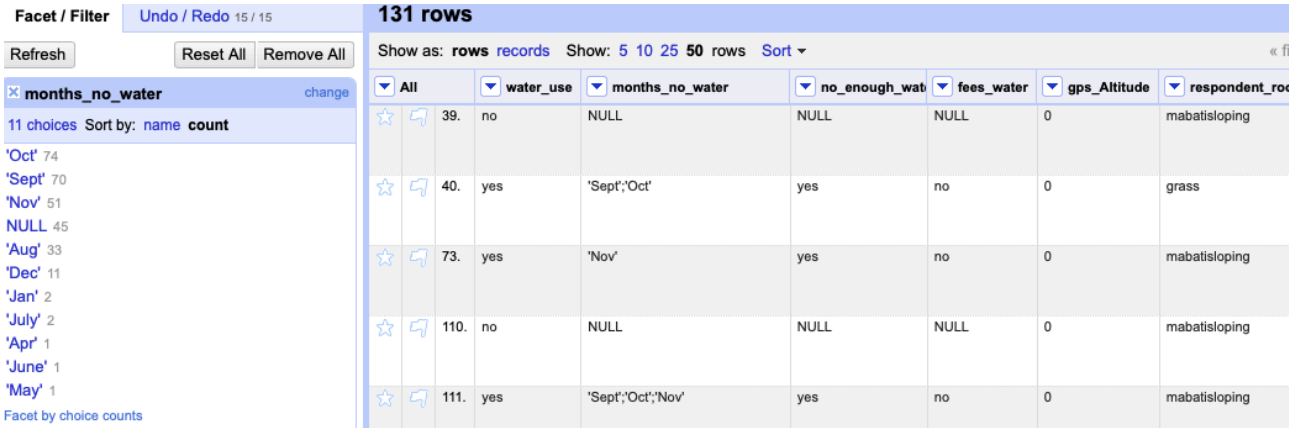
.

3. Fix the interviews dataset in OpenRefine enough to answer this question: "Which 2 month(s) are the most water-deprived/driest for the interviewed farmer households?"

The driest months are October and September.

4. Describe briefly the steps you took to achieve the answer to point 4.

We figured this out by putting forward all the data that had something to do with water (water_use, months_no_water_, no_enough_water, fees_water). We could then see that 'months_no_water' was measured in months, and therefore we took this data and sorted it to get the answers.



131 rows

Show as: rows records Show: 5 10 25 50 rows

months_no_water

11 choices Sort by: name count

'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

All	water_use	months_no_water	no_enough_wat	fees_water	gps_Altitude	respondent_roo
39.	no	NULL	NULL	NULL	0	mabatisloping
40.	yes	'Sept';'Oct'	yes	no	0	grass
73.	yes	'Nov'	yes	no	0	mabatisloping
110.	no	NULL	NULL	NULL	0	mabatisloping
111.	yes	'Sept';'Oct';'Nov'	yes	no	0	mabatisloping

We sorted it this way:

- Edit cells -> transform
  - value.replace("[","").replace("]","")
  - The replace function replces something. In this case we remove [ ] and replace it eith nothing
  - Afterwards we also used this function to remove the space: value.replace(" ","")
- We used the split-function to split the string in smaller pieces. In this case through semicolon ;
  - Facet → custom text facet.
  - In the expression we wrote: value.split(";")
- Then we got the answers in the left side and we sorted it by count!