Homework for Monday, 14th November 2022
Aarhus Universitet

W45: Open Refine

1. **Create a *tidy* spreadsheet/table listing the names of Danish monarchs with their birth- and death-date and duration of reign. They should be sortable by year of birth. Suitable source websites are for example here, but you can also use another source, provided you reference it. (Collaboration is welcome. Remember to attach this spreadsheet to Brightspace submission)**

   I've attached the Excel-spreadsheet to the submission, but here is also a Screenshot of the Spreadsheet:

	A	B	C	D	E	F
1	name	Birth_Year	death_Year	Start_Of_Reign	End_Of_Reign	Duration_Of_Reign
2	Gorm	908	958	N/A	N/A	N/A
3	Harald_1	N/A	978	N/A	N/A	N/A
4	Svend_1	N/A	1014	N/A	N/A	N/A
5	Harald_2	N/A	1018	1014	1018	4
6	Knud_1	995	1035	1018	1035	17
7	Hardeknud	1020	1042	1035	1042	7
8	Magnus_1	1024	1047	1042	1047	5
9	Svend_2	N/A	1076	1047	1074	27
10	Harald_3	N/A	1080	1074	1080	6
11	Knud_2	N/A	1086	1080	1086	6
12	Oluf_1	N/A	1095	1086	1095	9
13	Erik_1	1056	1103	1095	1103	8
14	Niels	N/A	1134	1104	1134	30
15	Erik_2	N/A	1137	1134	1137	3

   As you can see in the ""Birth_Year", "Start_Reign" and "End_Reign", there is data that is missing. I've used the data from https://www.kongehuset.dk/monarkiet-i-danmark/kongerakken as I find this website as being the most reliable. But I've also doublechecked with https://danmarkshistorien.dk/vis/materiale/kongeraekken, and they say the same thing. So instead of just leaving the cells blank, I've written "N/A", as an indicator of this missing data
   Also, the earliest king's birth-year (e.g., Gorm) is not 100% certain, s0 I've used one of the possible years (see the description of the kings from the Danish monarchy's website)
   To calculate the kings reign-duration, I've just said: = [cell 1] –[cell 2] the first two times, and then I've pulled the box down, in order to get the reign-duration for the other kings

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

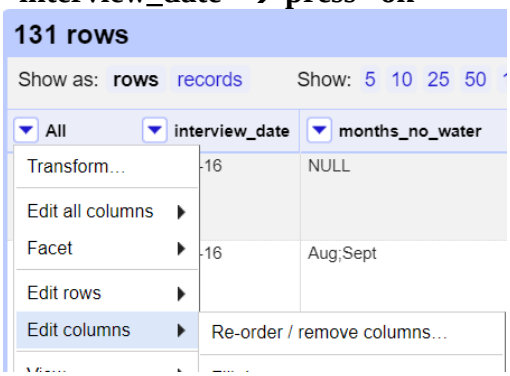
   No. If you are using a Spreadsheet with raw-unrefined data (e.g., the Danish monarchs) in OpenRefine, you are not altering the raw data. By using OpenRefine, you are using another platform, where you insert a "copy" of the dataset. You can sort and filter as much as you want in OpenRefine, but you will still have the original set on your computer (e.g., from Excel). This also applies to data inserted by a URL-link (see the optional task)

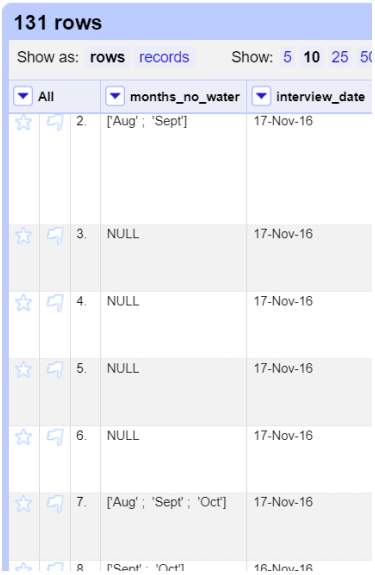
Homework for Monday, 14th November 2022
Aarhus Universitet

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?"**

The way I sorted and fixed the dataset in OpenRefine:

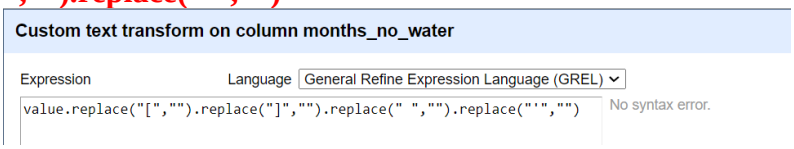
1. Inserted the data
2. Move the column "month_no_water" next to the interviews-date: **the arrow at "all" → edit columns → reorder/remove columns → drag the "months_(…)" next to the "interview_date" → press "ok"**



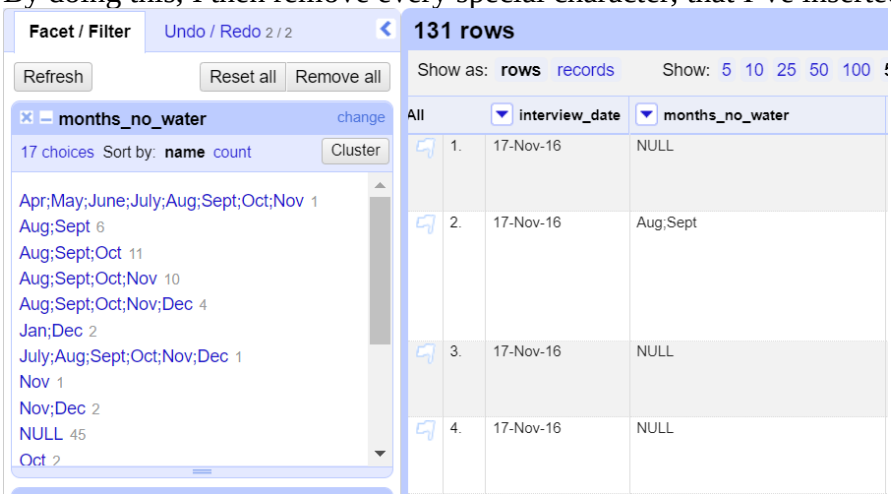


As you can see (picture right), the data is kind of mixed and with several kinds of data (here: months) in one cell, which is not what we want. So, I tidy up the data like this:

3. **Arrow (menu at "months[…]") → edit cells → transform**
4. After I've pressed the "transform…" function, I then write this expression in the bar:
   value.replace("[","").replace("]","").replace(" ","").replace("'","")

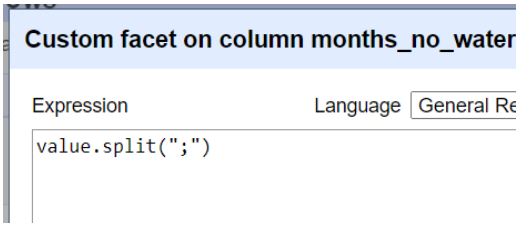


By doing this, I then remove every special character, that I've inserted in the expression



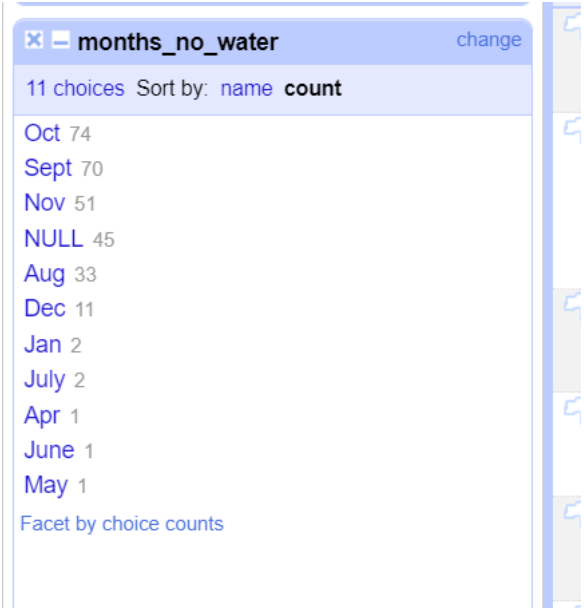
As you can see to the right, there is only **";"** left in the "months_no_water". But to answer the question: "which two months is the driest?", I have to replace the **";"**, and the way I do this is, by:

a. **Arrow → facets → custom text facets**:



After I've written this facet, I will then get this (right picture), but to answer the question, I then sort the data by "count", which then makes me able to answer the question.

- the two most dry months reported by the interviewed families are: **October and September**.



4. **OPTIONAL Real-Data-Challenge: What are the 10 most frequent occupations (erhverv) among unmarried men and women in 1801 Aarhus? (hint: some expert judgement interpretation is necessary)**

To get the two results (pictures) I did the following:

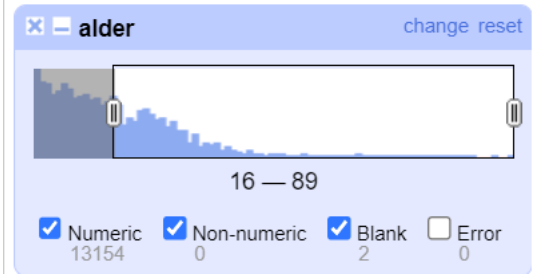
1. Inseated the data from the website, by using the URL-option :

2. Reorder the columns: erhverv, køn, alder, civiltilstand, so the four of them are closer to the beginning
3. Make a "text facet" in the "koen [køn]" column:

← this makes me able to see resp. male or female

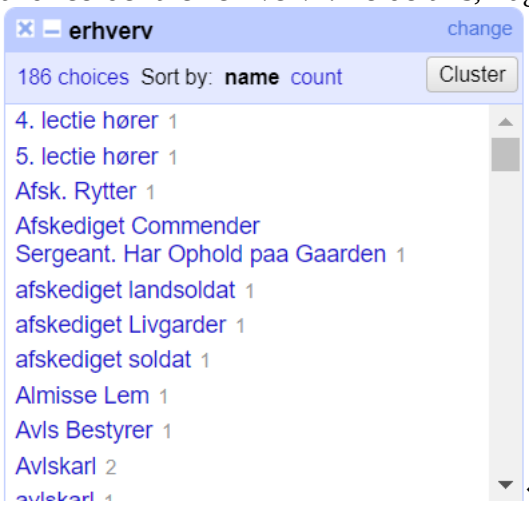
4. Select "mand" [male], which then only show me the male data
5. I then make another "text facet" but this time in the "civilstatus" column. Then I get numerous results, but I choose the "ugift" [unmarried] option. I then only get the unmarried, male data (see picture below)

6. I then make another facet, but this time with numbers, under the column "alder" [age], which then makes me able to select a specific age-period, so I don't get the boys under e.g., 16



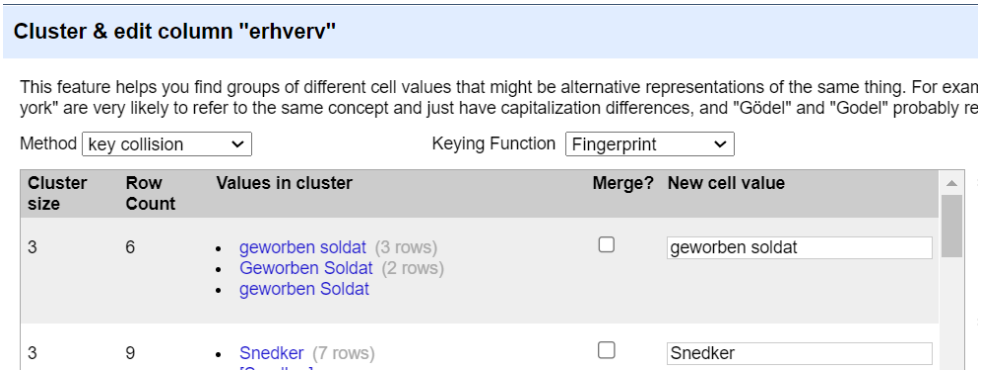
← by doing this, my workspace will only include the unmarried men, between 16-89 y/a. If you look at the facet (above), you can see that after the age on around 40, most men are married, so we wont be able to find that many unmarried "old" men

7. In order to find out the 10 most occupied jobs among unmarried men, we have to filter and reorder the "erhverv". To do this, I again make a text facet (not a custom one):

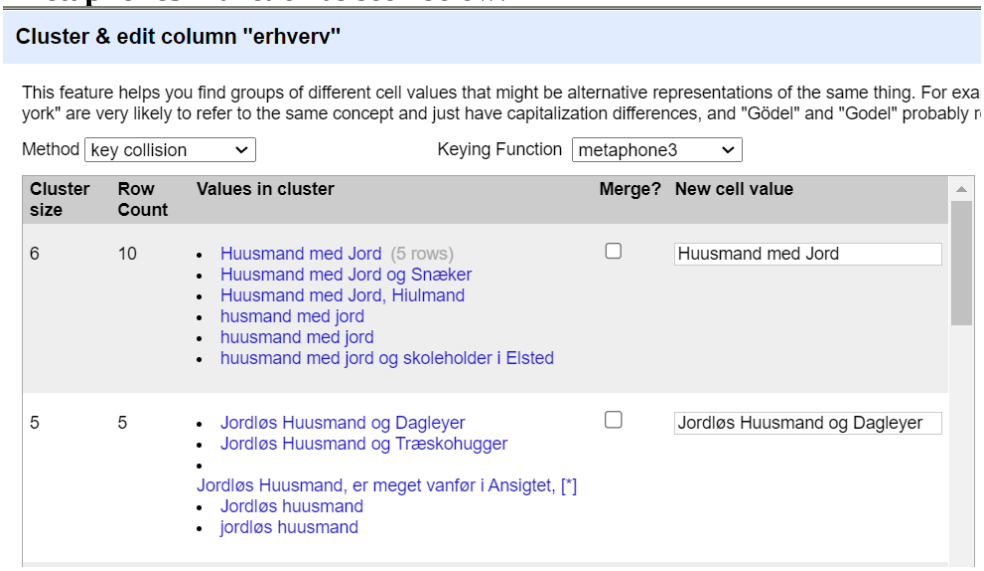


← but in order to see the most jobs, I then use the "cluster" function:

Homework for Monday, 14th November 2022
Aarhus Universitet



← I firstly use the "**fingerprint**" function and merge the different values. The I use the "**metaphone3**" function as seen below:



By using this function, I will be able to merge the different types of "erhverv" (spelling-wise), so they will count as "one"
I know this way maybe isn't the correct one, as you have to repeat this process a couple of times, but it gets the job done. By repeating this process, you are able to "catch" the spelling errors and merging them with the correct ones. After repeating this around 3 times, I get the results as seen below:

← I can then say that the 10 most common types of jobs among the unmarried men between 16-89, is (see picture).

Homework for Monday, 14th November 2022
Aarhus Universitet

8. When it comes to the unmarried women, I've done the same thing, but of course by choosing the "kvinde" option under the column-facet "køn".
   The results I've gotten is: →

I'm guessing this way I've done it isn't the correct way, and of course you have to keep in mind that there (at that time) was different ways of describing your job, and that there is some spelling errors, that my way of clustering-it didn't fix. But I'm about 90% sure in my results, when it comes to the 10 most common types of jobs:

Men: Soldier, farmer, craftsmen (more physical-hard jobs)
Women: maiden, spinning-lady (more "quiet-lady-like jobs)

I don't know if its useful for you, but I've also attached the "redo/undo" history in another PDF-document. I don't know how to read them, but maybe you do 😊