Anton Drasbæk, AU682983

2:W35: Open Refine

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

See file: danish_monarchs.csv on the author's GitHub repository (https://github.com/Digital-Methods-HASS/au682983_Schioenning_AntonDrasbaek)

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

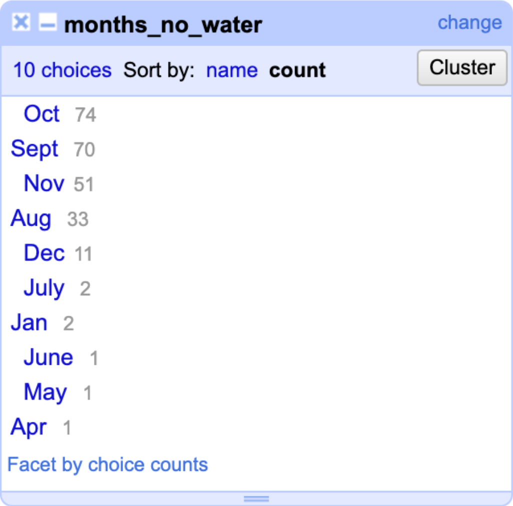
It does not. When sorting, it restructures the data so that the order of observations are changed. When filtering, it removes some observations based on a logical expression. However, these are NOT removed from the raw data itself.

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?"

Firstly, we identify the variable "months_no_water" which presumably indicates which months a particular household had no water. We make the assumption that the most frequently mentioned months under this variable are the most water-deprived months by interviewed farmer households.

In OpenRefine, we filter the "months_no_water" column to exclude all "NULL" observations. We then make a transform which removes everything but the month_name and semi-colon separator (value.replace("'", "").replace("[","").replace("]","")). Next, we use the "Split multi-valued cells" function to split up the month_no_water column. Finally, we cluster the values as some months include and space and others don't, which mean they were not grouped.

Finally, we conclude that October (74 households) and September (70 households) were the months reported as the most water-deprived by interviewed farmer households:



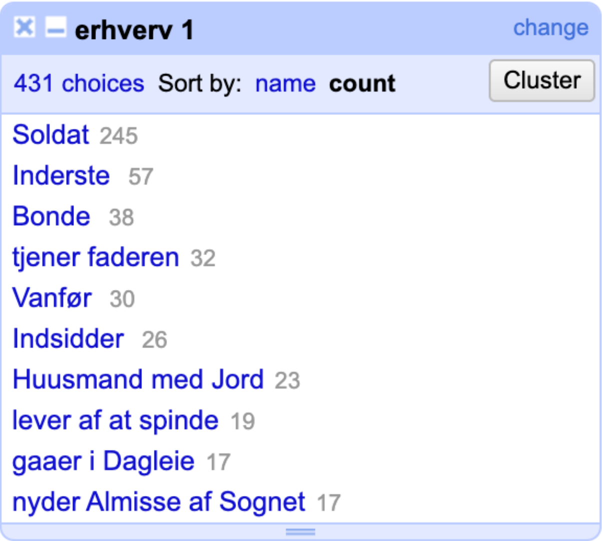
Oct	74
Sept	70
Nov	51
Aug	33
Dec	11
July	2
Jan	2
June	1
May	1
Apr	1

4. 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, look at the HISCO classification "Historical International Standard of Classification of Occupations" on Dataverse if ambitious)

On OpenRefine, I firstly applied a filter to the “Civilstatus” column to only include those who were “Ugift”, as we are only concerned with unmarried people. Then I applied a filter to “Erhverv” that removed all empty columns. Furthermore, “Erhverv” was then split based on the regular expression “on ,|\bog”. This splits the column on all commas and the word “og”. From here, we have three “Erhverv” columns. We make the assumption that the first column is the person’s primary occupation and thus we only work with that going forward.

A text facet is then applied to “Erhverv 1”. As many occupations seems to be the same but with variation in spelling, we do clustering. First, key clustering is applied and afterwards nearest neighbor is almost used. Finally, anything including “Soldat” is grouped into one group as we assume being a soldier is an general occupation.

This leaves us with the final 10 most frequent occupations for unmarried men and women in 1801 Aarhus:



Soldat	245
Inderste	57
Bonde	38
tjener faderen	32
Vanfør	30
Indsidder	26
Huusmand med Jord	23
lever af at spinde	19
gaaer i Dagleie	17
nyder Almisse af Sognet	17