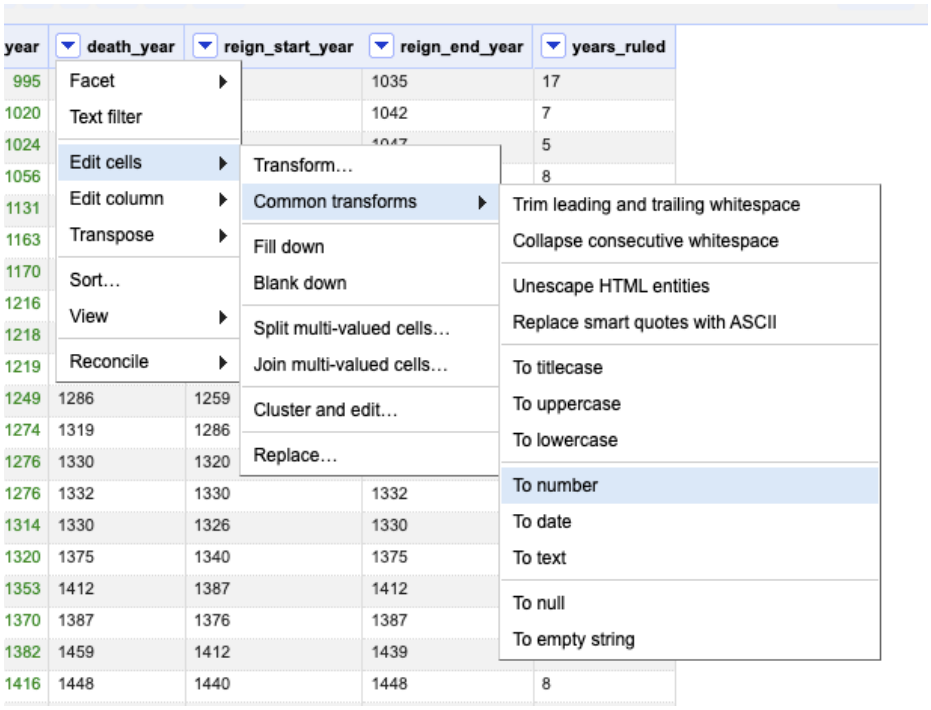
Homework 2 week 35: Open Refine

Task 1: create a tidy spreadsheet and sort monarchs by borth year

I created a spreadsheet listing the names of the Danish monarchs with their birth- and death-date and start and end year of reign. I use the data from danmarkshistorien.dk and kongehuset.dk and put it into an excel spreadsheet called `monarchs.csv` (can be found in the data_hw2 repository on GitHub: https://github.com/Emma-Marie/final_project/tree/main/final_project). I named the columns `danish_monarchs`, `birth_date`, `death_date`, `reign_start_year` and `reign_end_year` to sort the data. The birth date of many of the ancient kings were unknown, and I used the name NULL to mark the missing data. The fact that the year of birth has its own column makes the data sortable by year of birth.

I pulled the data set into OpenRefine, and converted the years into numbers clicking on "edit cells" > "Common transforms" > "To number"



I sorted the monarchs by year of birth by clicking on the dropdown menu at the column named `birth_year` and choosing "sort", "numbers" and "smallest first".

All		danish_monarchs	birth_year	death_year	reign_start_year	reign_end_year	Column
5.		Knud (2.) den Store	995	1035	1018	1035	
6.		Hardeknud (Knud 3.)	1020	1042	1035	1042	
7.		Magnus (1.) den Gode	1024	1047	1042	1047	
12.		Erik (1.) Ejegod	1056	1103	1095	1103	
18.		Valdemar (1.) den Store	1131	1182	1157	1182	
19.		Knud 6.	1163	1202	1182	1202	
20.		Valdemar (2.) Sejr	1170	1241	1202	1241	
21.		Erik (4.) Plovpenning	1216	1250	1241	1250	
22.		Abel	1218	1252	1250	1252	
23.		Christoffer 1.	1219	1259	1252	1259	
24.		Erik (5.) Klipping	1249	1286	1259	1286	
25.		Erik (6.) Menved	1274	1319	1286	1319	
26.		Christoffer 2.	1276	1330	1320	1330	
28.		Christoffer 2.	1276	1332	1330	1332	
27.		Valdemar (3.) Eriksen	1314	1330	1326	1330	
29.		Valdemar (4.) Atterdag	1320	1375	1340	1375	
31.		Margrete 1.	1353	1412	1387	1412	
30.		Oluf 2.	1370	1387	1376	1387	
32.		Erik (7.) af Pommern	1382	1459	1412	1439	
33.		Christoffer 3. af Bayern	1416	1448	1440	1448	
34.		Christian 1.	1426	1481	1448	1481	

References:

Kongehuset.dk, ”Kongerækken”, https://www.kongehuset.dk/monarkiet-i-danmark/kongerakken, visited September 2022

Danmarkshistorien.dk, “Kongerækken, ca. 950-“, https://danmarkshistorien.dk/vis/materiale/kongeraekken, last edited October 7 2021, visited September 2022

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

OpenRefine doesn’t alter the raw data during sorting and filtering. The raw data is always to be found in the table, while the clustering, faceting, and editing of the data is to be seen in the small windows in the left side of the screen. Furthermore, it is always possible to go back to earlier versions of the dataset using the undo/redo function in the upper left corner of the window.

Task 3: Fixing the interviews data set in OpenRefine to answer the question: "Which two months are reported as the most water-deprived/driest by the interviewed farmer households?"

Firstly, I pull the attached dataset into a new OpenRefine project. To make it easier for myself, I find the column called `months_no_water`, click on the dropdown menu, and choose “edit column” and “move column to beginning” to move it to the beginning of the table. Then I make a text facet by clicking on the dropdown menu and clicking on “facet” and then on “text facet”.

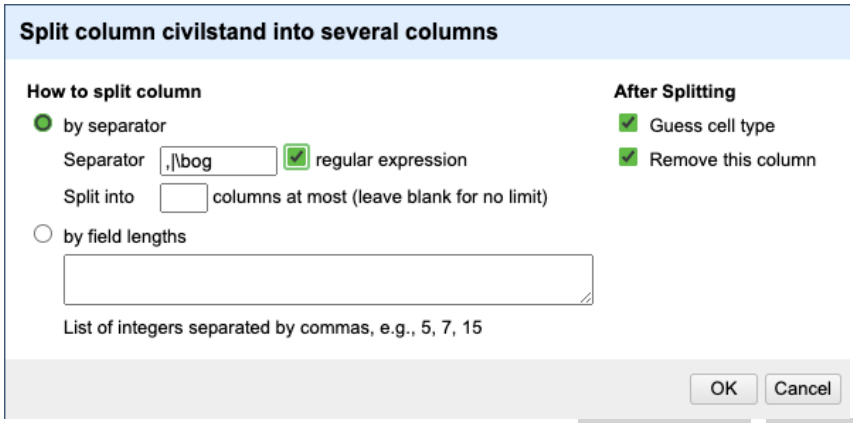
Some of the interviewed households have reported several months to be the driest. To find the two driest months, I must make OpenRefine recognize the months separately. I need to remove the single quotation marks, the square brackets, and the spaces, and seperate the observations by the semicolons. In the dropdown menu I choose “facet” and “costume text facet” and type the expression `value.replace("[", "").replace("]", "").replace(" ", "").replace("'", "").split(“;”)`. The `value.replace` command replaces the sign in the first set of double quotation marks with the sign in the second set of double quotation marks. To remove the signs, I don’t write anything in the second set of quotation marks. The command `value.split` splits the observations by the sign, which you write in the double quotation marks. In the new text facet popping up the months are listed separately. I click on “count” to list them by frequency:

October and September are the two months, which are reported the most water-deprived or driest by the interviewed farmer households.

Task 4: answer the question: "What are the 10 most frequent occupations (erhverv) among unmarried men and women in 1801 Aarhus?"

I pull the data from the attached csv file into OpenRefine and move the columns `erhverv` (occupation) and `civilstand` (marital status) to the beginning of the table for the sake of convenience.

Some of the rows in the `erhverv` column have several occupations in one cell, and I am interested in splitting the column into several columns. I assume, that the first mentioned occupation is the main occupation, and I therefore ignore the secondary occupations in my analysis. I want to split the occupations written in the same cell by the "og" and the commas. I click on the dropdown menu and choose "edit column" and "split into several columns…", and type the regular expression `,|\bog` to move the occupations which comes after a comma or after an "og" to their own separate columns.



Now I have five occupation collumns called `erhverv 1`, `erhverv 2`, `erhverv 3`, `erhverv 4` og `erhverv 5`. Only the `erhverv 1` column is relevant for the task.

I make the `erhverv 1` column into a text facet and use the cluster function to make sure that miss-spellings or difference in spelling is gathered under the same names. Some of the last clustering I did manually by scrolling through the text facet and editing the misspellings and different spellings of the same occupation.

To only see the occupation of the unmarried, I make a textfilter on the `civilstand` column in which I type "ugift" (unmarried). Now the text facet from the column `erhverv 1` only lists the occupations belonging to unmarried persons. The ten most frequent occupations are (ordered after frequency) national soldier, soldier by the 1st Jutland infantry regiment, lives

for rent on a farm (inderste), country soldier, female servant, invalid, weaver, male servant, apprentice, and spinner.