**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”**Graphical user interface, table

Description automatically generated**

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”.

Table

Description automatically generated

**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:

**Graphical user interface, text, application

Description automatically generated**

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.

Graphical user interface, text, application

Description automatically generatedSome 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.

Graphical user interface, application

Description automatically generated