Homework W37

1. **Create a \*tidy\* spreadsheet/table listing the names of Danish monarchs with their birth- and death-date and start and end of their reign. They should be sortable by year of birth. Suitable source website is for example** [**here**](https://kongehuset.dk/monarkiet-i-danmark/kongerakken)**, but you can also use another source, provided you reference it. (Collaboration is welcome. Remember to attach this spreadsheet to Brightspace submission)**

Knud den Store (995-1035) is considered here as the second Knud. The first Knud was Knud (1.) Hardegon (died before 935), also called Hardeknud Svendsen, father of Gorm den Gamle(?-958).

Gorm den Gamle’s year of birth is estimated to year 908 and year of birth, 958

Harald Blåtand’s year of death is estimated to year 987

We made the spreadsheet tidy by making sure that we used the different principles when working with data in spreadsheets. For instance we made sure to only use one variable in each column, and we only used one variable when describing data that could not be found in the dataset.

Furthermore, when dates and other data was not available to us we wrote NA (not applicable) to make it clear to the reader that we know that the cell dosen’t contain data.

Link to spreadsheet:

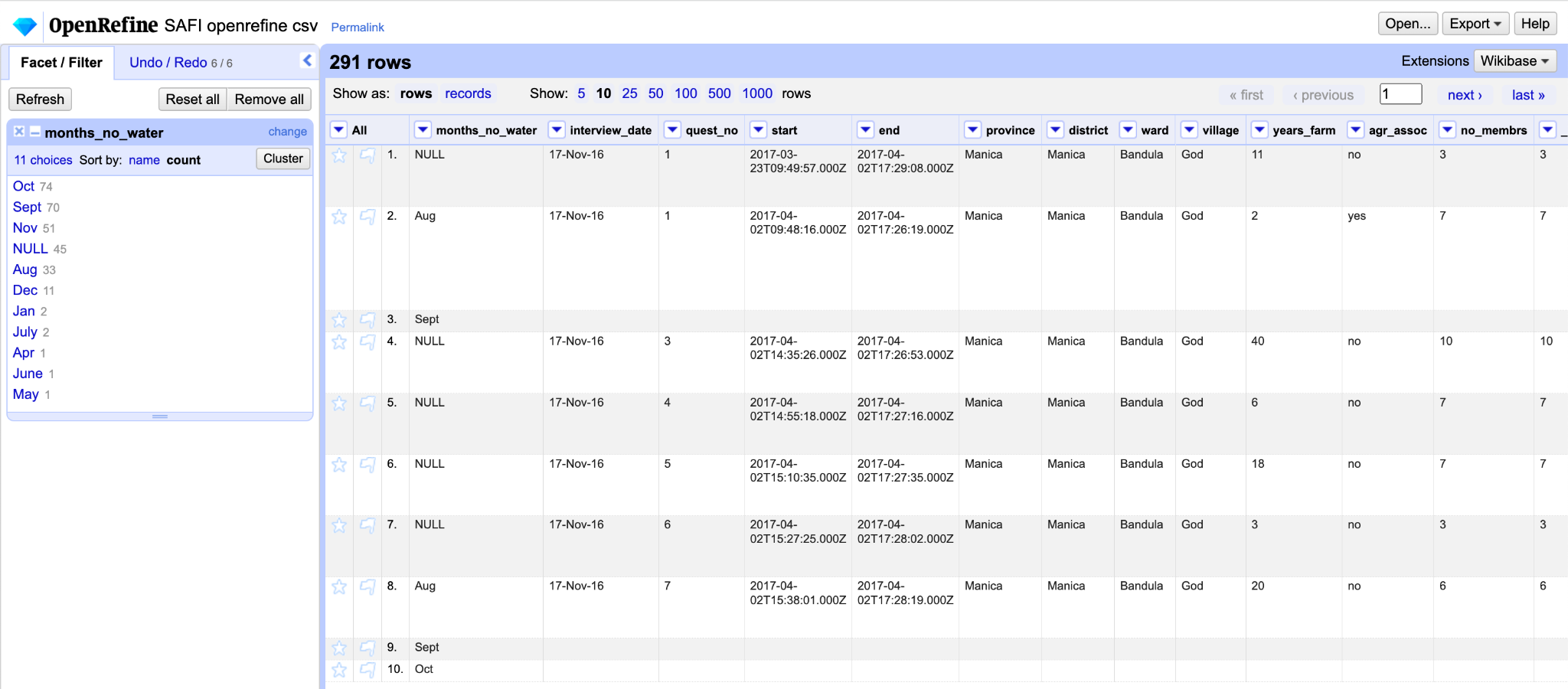
<https://docs.google.com/spreadsheets/d/1lQ9IExlS9EuHBU_Edc5_f8uuoa3WfEsj_X8DEJCHqs8/edit#gid=0>

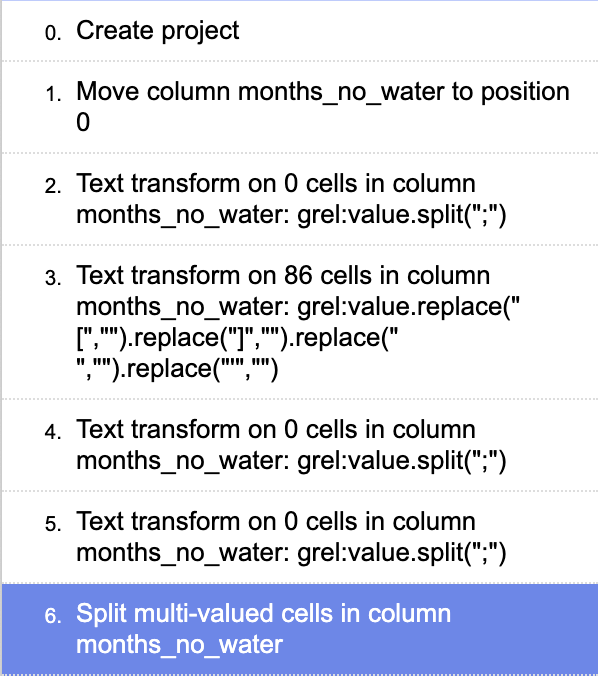
1. **Does OpenRefine alter the raw data during sorting and filtering?**

Openrefine does not alter the raw data. Openrefine generates a code in which you are able to see your modifications. You can always find the different steps and alterations under the function called “undo/redo”

It can be useful to extract to save with the data in order to reproduce the “experiment”

1. **Fix the** [**interviews dataset**](https://ndownloader.figshare.com/files/11502815) **in OpenRefine enough to answer this question: "Which two months are reported as the most water-deprived/driest by the interviewed farmer households?"**



<https://docs.google.com/document/d/1txvsn3FxXAOL72md-NxWO38BoTXf-CtgNBJi6ky9uXY/edit?addon_store>

The two months which are reported to be the most water-deprived/driest by the interviewed farmer households are October (80) and September (70).

We fixed the dataset and found the answer by:

* Moving the column called “month\_no\_water” to the beginning of the spreadsheet in OpenRefine.
* Then we transformed the text by using GREL, where we removed the square brackets, space and apostrophe. The text in the cells were now separated by semicolons instead.
* Afterwards, we used the function “split multi-valued cells” under “edit cells”. This made it possible to split the months into separate values.

1. **OPTIONAL Real-Data-Challenge: What are the 10 most frequent occupations (erhverv) among unmarried men and women in** [**1801 Aarhus**](https://raw.githubusercontent.com/aarhusstadsarkiv/datasets/master/censuses/1801/census-1801-normalized.csv)**? (hint: some expert judgement interpretation is necessary. As an inspiration,** [**check out this chapter Making a living outside marriage from the**](https://academic.oup.com/book/1458/chapter/140849368) **Swedish Gender and Work project of Maria Agren)**

First I uploaded the .cvs file in OpenRefine, found the column with “civilstand” and used the “text facet” function and then included “ugift”



A lot of clustering in order to make the search case- and spelling insensitive.

Some of the cells are multi-valued, so I have “split multi-valued cells” with the separator “ og”

The same with “ samt “ and “, “

Afterwards I’ve done further clustering.

I’ve chosen to count every type of “soldat” under the label “Soldat”

Hereafter I used “text facet” on the “erhverv” column and chose “sort by count”. In order to show the most frequent matches.



Modifications and clusters:

<https://docs.google.com/document/d/1AqG43JzZdkLwaOpQZiVc4X2d-QDWneA4eV7Z0lw3oHc/edit>