# 2:W35: Open Refine

Upload your answers to these questions:

## 2.1) Danish monarchs, create spreadsheet

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 ATTACHED FILE monarchs.csv IN THE FOLDER hw_w5_2*

## 2.2) Does OpenRefine alter the raw data during sorting and filtering?

*Not during sorting and filtering. However, when the spreadsheet is uploaded to OpenRefine, the data types changes. They can quickly be transformed into a different format by clicking on the column arrow > Edit cells > Common transforms.*

## 2.3) Fixing the interviews dataset and look for water-deprived months

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?"

I'll use the variable *months_no_water* to answer the question: *Which two motnhs are reported as the most water-deprived/driest by the interviewed farmer households?"*.

I go to the column menu, choose the menu item « Split into several columns... »

7. 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)

First, I want to split the cells that contains 2 months into separate cells. I see that the cell contains hard brackets, the symbol ' and semicolon, and the semicolon is what separates the monts. I therefore split the multi-valued cells **by separator ;**.

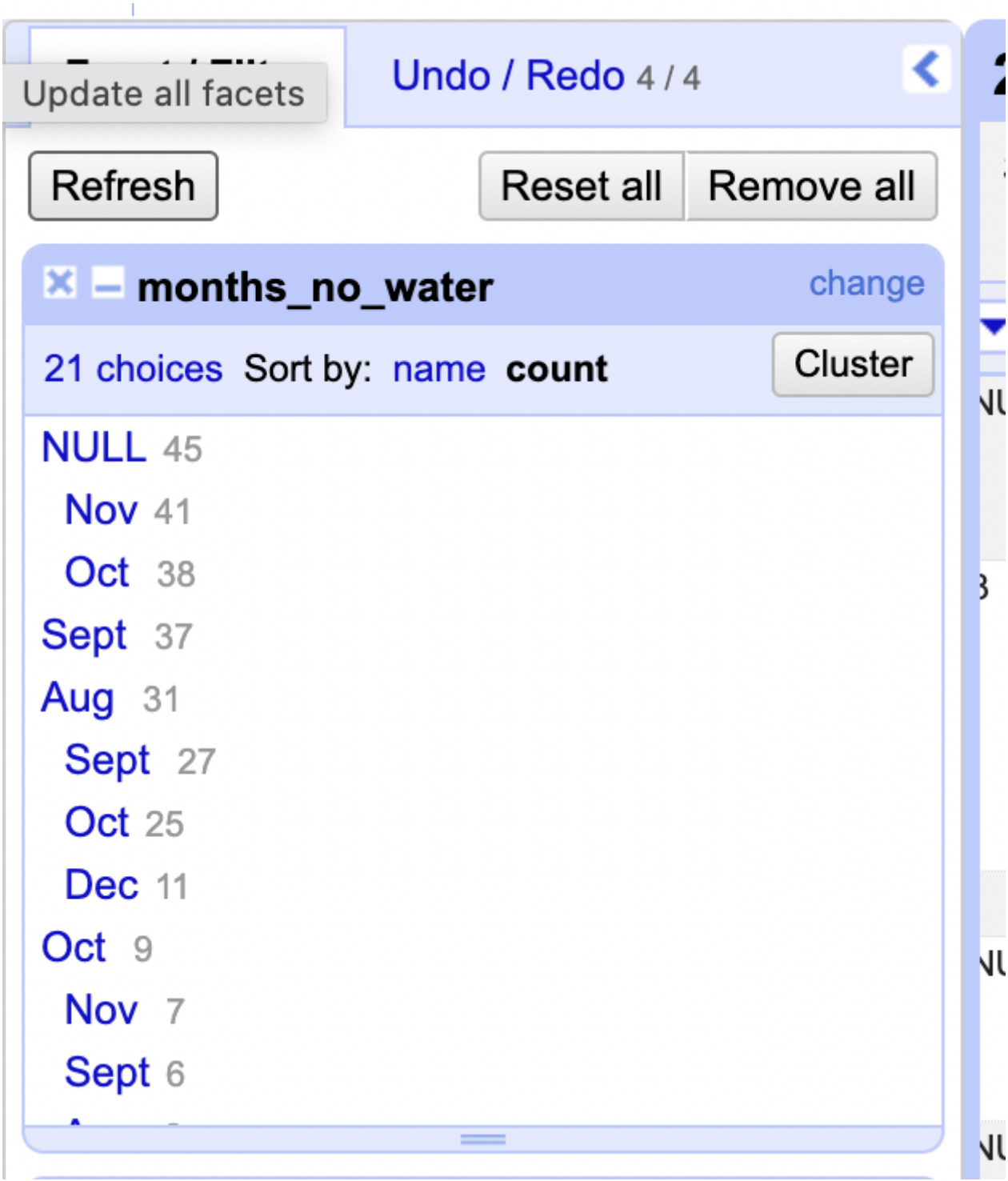
I here click the following:

- the small arrow next to the colum **months_no_water**
- **Edit cells**
- **Split multi-valued cells**..
- then I fill ; into the **Separator** field and click **OK**.

I now want to get rid of all the special characters. I do that by following these steps:

- Edit Cells
- Transform
- To remove all left square backets and replace with nothing: ```value.replace("[", "")
- To remove all right square backets and replace with nothing: `value.replace("]", "")`
- To remove all ' signs I write: `value.replace("]","")`
- I then go to the left pane and **Sort by count**.

*I see that the two months reported as the most water-deprived/driest are November (count = 41) and October (count = 38).*




## 2.4) 10 most frequent occupations

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)

Data: *https://raw.githubusercontent.com/aarhusstadsarkiv/datasets/master/censuses/1801/census-1801-normalized.csv*

HISCO classification: *https://github.com/cedarfoundation/hisco*

I first load the data set into OpenRefine by copy pasting the link.

I have to do some cleaning since some cells with **erhverv** contains more than one value. I therefore split by `og`, `´`

I want to filter by **civilstand = ugift** and then group by men and women. I'll use **Facet** which alllows me to group and also to filter the data by these values.

First I'll do a **Text Facet** which will group by identical text values in a specific column and then list unique values with the number of records it appears in.

So first, I scroll over the `civilstand` column. Then I click the down arrow and choose `Facet` > `Text facet`. In the left panel, I now see a box with every unique value in the `civilstand` column along with a number representing how many times that value occurs in the column. I click `ugift` to exclude all other categories.


I then do the same with `koen`. So first, I scroll over the `koen` column. Then I click the down arrow and choose `Facet` > `Text facet`. In the left panel, I now see a box with every unique value in the `koen` column along with a number representing how many times that value occurs in the column. I click `kvinde` to first have a look at the most frequent occupations for women.

I then do the same with the `erhverv` column. Then I click `Sort by: count` to see each count for each occupation for the women (because I selected **kvinde**). I then klick on **mand** in the **koen** pane and compare the counts.

| Gender | Female | Male |
| --- | --- | --- |

| Gender | Female | Male |
| --- | --- | --- |
| Occupations sorted by count |  |  |



**OBS: Scrolling down the facet panes, it becomes evident that the same occupations occur multiple times (as separate occupations) due to differences in spelling**. The duplicates that I could choose to get rid of by combining them into one category are among others:

- different types of *soldier*
- *stuepige*/*Stuepige*
- *vanvittig/Vanvittig*
- *gaardbeboer/gårdbeboer*
- _bonde/Bonde

The different categories can be searched for and replaced using RegEx (for instance soldier category `([\w-]+)(oldat)$`), but I just press in the left "erhverv" pane and edit manually. I see that the most frequent occupations for unmarried men is **soldat** (some sort of soldier) and the most frequent occupations (erhverv) among unmarried men and women in 1801 is **Tjenestepige**.