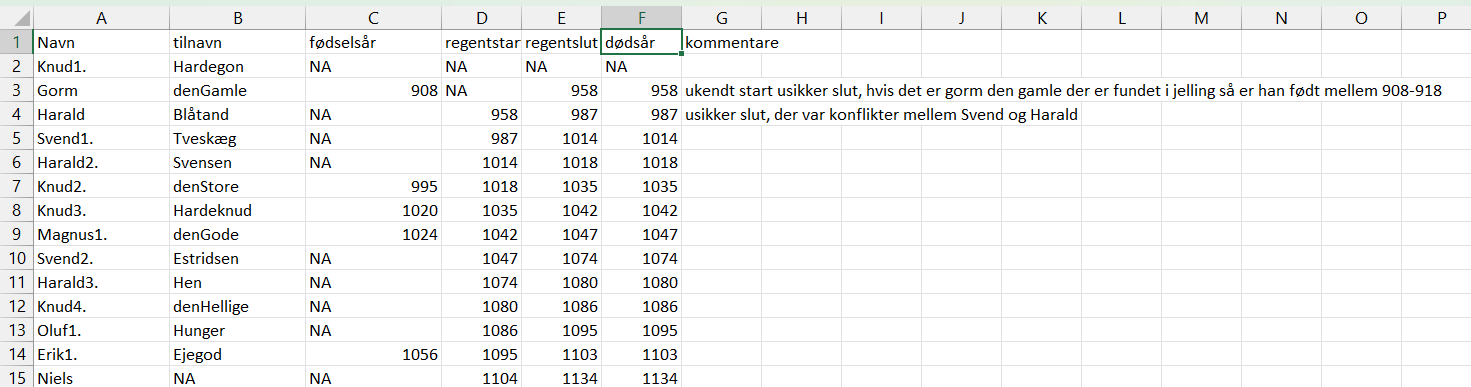
Assignment 2

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)***

In the spreadsheet that I made I choose to not have any spaces between the word and wrote all in one because I know myself well enough to know that I would mess it up with the underscore and then I set it up in birthyear, beginning-of-reign-year, end-of-reign-year and then death-year, which usually overlaps with the end of reign year.

I also chose to have a separate column of comments because there is some information that can be important depending on what you use the list for.



I have written the entire line of regents from Hardegon (Gorm den Gamles father) to Queen Magrethe in a spreadsheet which I uploaded to OpenRefine to see what happens when I try to sort it by birthyear[[1]](#footnote-1).

At first, I click on the arrow at fødselsår (red arrow), then sort and then sort again, because it is a year and not a date, means that I can’t use the timeline function in facets





Then I choose numbers (blue arrow) and smallest first (green arrow) and press okay.

Et billede, der indeholder tekst

Automatisk genereret beskrivelse



Then I run into the following problem as with a lot of the older kings we simply don’t know when they were born which leads the bottom of the list to look like this.

Et billede, der indeholder bord

Automatisk genereret beskrivelse

A solution to this could be to sort by when their reign began, but here there are a few we don’t know, but could maybe do an approxiamation of so it is at least in the correct order, which is easy enough to do with kings as we know which order it is supposed to be in, but would make it more difficult if we were working with a list of normal peoples birth year.

<http://127.0.0.1:3333/project?project=2589494417311>

[

{

"op": "core/text-transform",

"engineConfig": {

"facets": [],

"mode": "row-based"

},

"columnName": "fødselsår",

"expression": "value.toDate()",

"onError": "keep-original",

"repeat": false,

"repeatCount": 10,

"description": "Text transform on cells in column fødselsår using expression value.toDate()"

}

]

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

OpenRefine does not alter the raw data as it is a file or an URL, that is put into the OpenRefine system, not the acutal data. You can then afterwards download the changed data, but still have the raw data. Even though it isn’t the raw data that is changed.

In the case of the line of regents the lack of information about birth year becomes a problem when trying to sort the information by birth year, and also as it can be seen in the comments I have in the side of the document, a lot of history can be unsure, Gorm den Gamle was born between 908-918 if the man found buried in Jellinge is in fact him and all this uncertainty is not shown, when you just put a spreadsheet with numbers or single words into OpenRefine.

3) ***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?"***

First I move the column of months\_no\_water to the first column in OpenRefine, you do that by pressing the little arrow at the top of the column and then edit column and then move column to beginning.

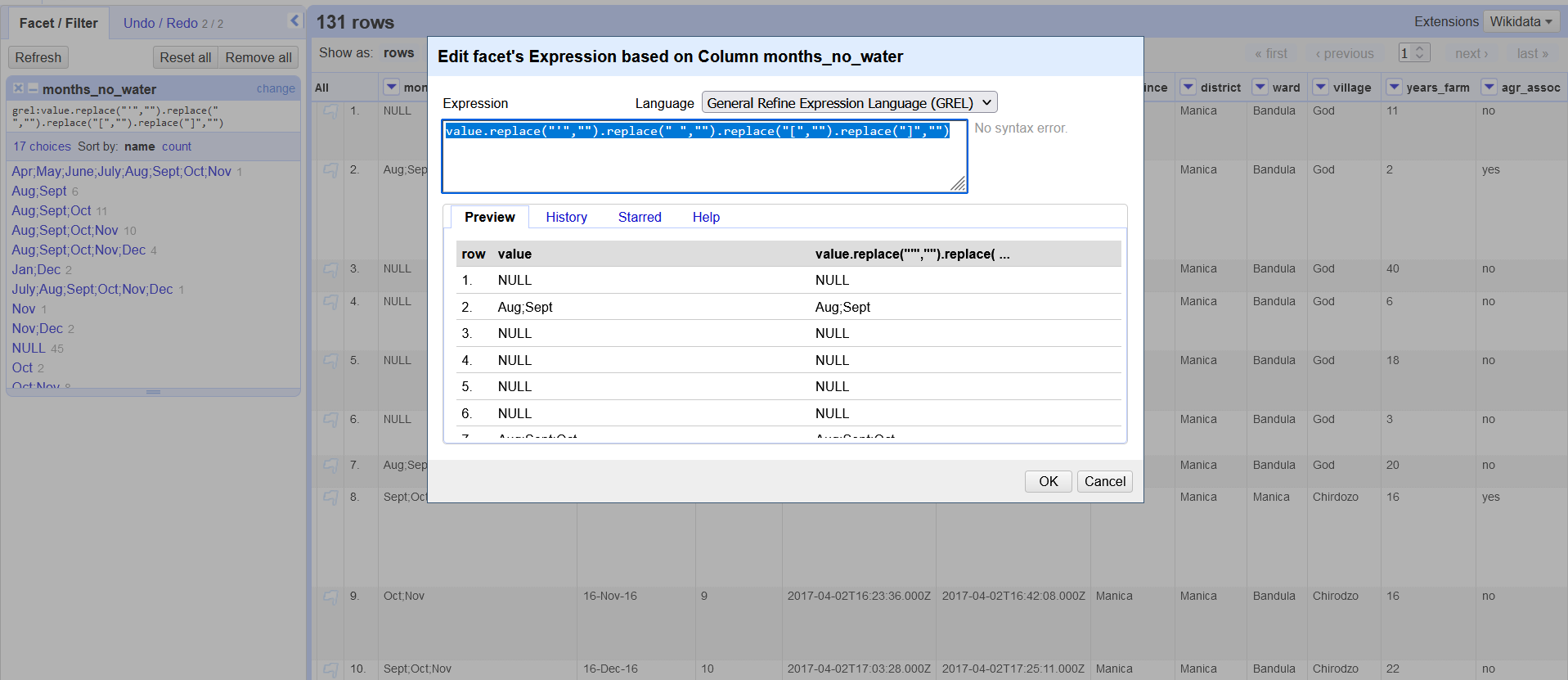
To figure out which months are the driest I have to clean op the data in the column as there is up to 3 different months in on field.

The first thing I do is clean up all unnecessary characters which I do by clicking the little arrow, then edit cells and then transform.

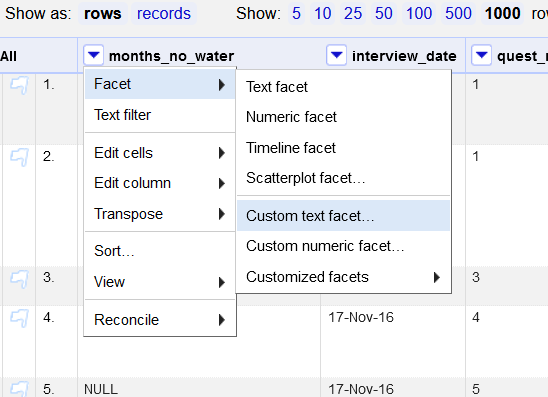
Et billede, der indeholder bord

Automatisk genereret beskrivelse

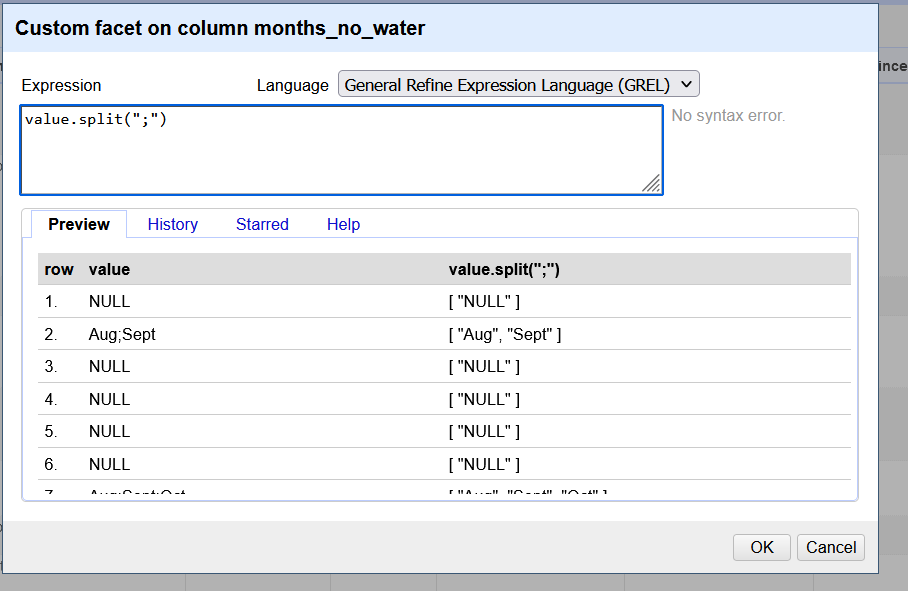
When I chosen transform a popup window shows up in that pop up window I write: value.replace("'","").replace(" ","").replace("[","").replace("]",""), the thing before the comma being replaced with what is after the comma, which in this case is qoutationsmarks, space, and brackets, I keep the semicolon as I need it in the next step



The next step is to split up the different month. Again, I first choose the little arrow again, then facet and then custom text facet as shown below



Then I write value.split(“;”) which show that I want to separate the words in the column by the semicolon



Then there is a box at the side of the field, where you can different months as it is now it is not in order of how often they are mentioned

Et billede, der indeholder tekst

Automatisk genereret beskrivelse

That happens when you click count, then you see which month is mentioned most often which in this case is October and September

Et billede, der indeholder tekst

Automatisk genereret beskrivelse

[

{

"op": "core/column-move",

"columnName": "months\_no\_water",

"index": 0,

"description": "Move column months\_no\_water to position 0"

},

{

"op": "core/text-transform",

"engineConfig": {

"facets": [],

"mode": "row-based"

},

"columnName": "months\_no\_water",

"expression": "grel:value.replace(\"'\",\"\").replace(\" \",\"\").replace(\"[\",\"\").replace(\"]\",\"\")",

"onError": "keep-original",

"repeat": false,

"repeatCount": 10,

"description": "Text transform on cells in column months\_no\_water using expression grel:value.replace(\"'\",\"\").replace(\" \",\"\").replace(\"[\",\"\").replace(\"]\",\"\")"

}

]

4)

***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" \t "_blank)? (hint: some expert judgement interpretation is necessary)***

First thing I do I move the column with Erhverv to the beginning of the field so it is easier to for me to see when I work with it

Et billede, der indeholder bord

Automatisk genereret beskrivelse

Then is choose that I only wish to see the unmarried people by choosing civilstand and then facet and then choosing text facetEt billede, der indeholder bord

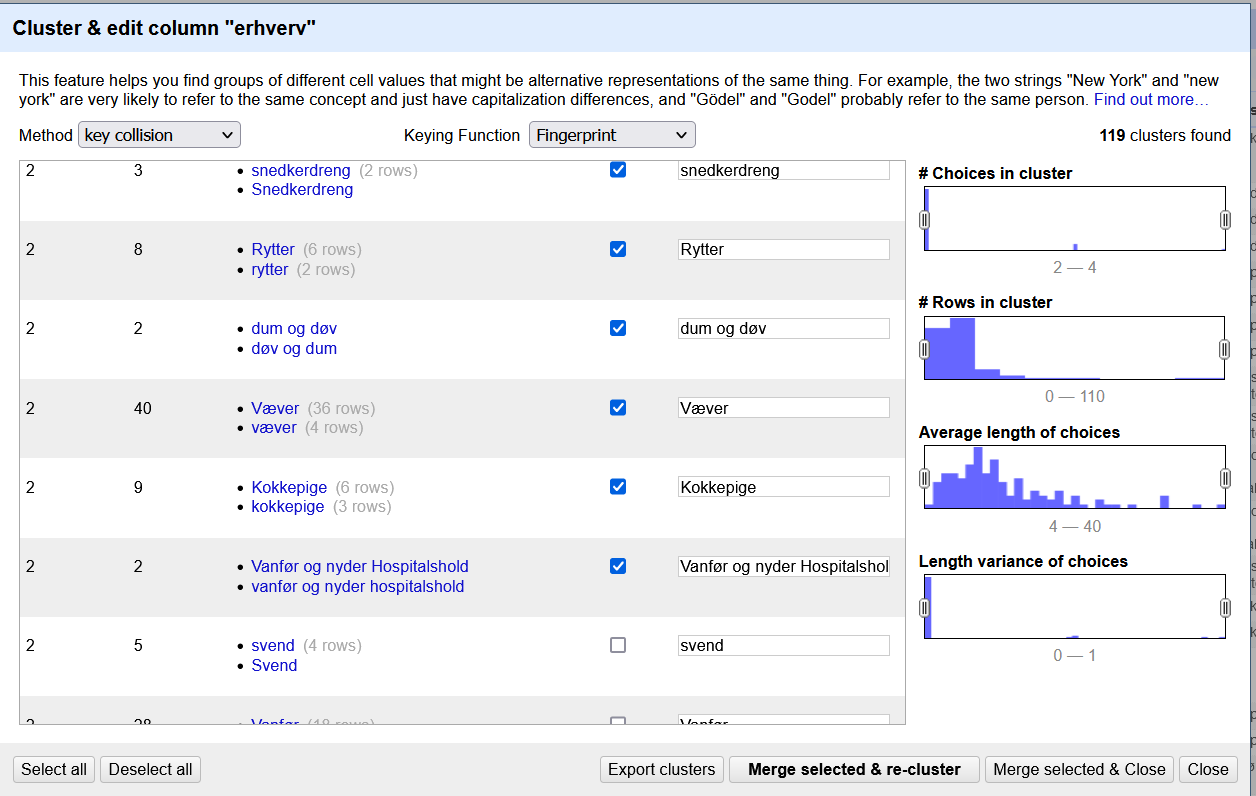
Automatisk genereret beskrivelse

Choosing textfacet I get a block looking by the one below, where I then pick ugift

Et billede, der indeholder tekst

Automatisk genereret beskrivelse

Then I have to cluster quite a few different jobs, an example of which is seen below, do that by pressing cluster and then using fingerprint and cologne-phonetic. This requires quite a few judgement calls and I was quite conservative in what I chose to cluster were it was mainly different spelling of the same words. The list of what I clustered can be seen in the code in the other file.



After I have clustered the different words I get the result seen below



Here it is worth noting that I could have clustered the three different kinds of soldiers, but I am frankly unsure of whether they are actually doing the same thing, as in the beginning of the 19th century nationalsoldater might mean people from Denmark or from Schleswig or Holstein or a combination thereof and that could be an important detail.

The code to this is quite long and is saved in a separate file called kode til opgave 4.

<http://127.0.0.1:3333/project?project=1890068156230>

1. Source for line of regents: <https://danmarkshistorien.dk/vis/materiale/kongeraekken>, note to Hardengon: you find the information to Hardegon if you click the footnote to Knud den Store [↑](#footnote-ref-1)