Joshua Hansen
201905780
BA.INF – Cultural Data Science
201905780@post.au.dk

Cultural Data Science assignment #2

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)*

For the first task, we took the information of this website and copy/pasted into a spreadsheet. We know had to split the information of the column into fitting and tidy colums. We used a split() method within Google Sheets and split it into the colums: “Regent_name” , “Reign_start” and “Reign_end” , where we put NA as the value for missing information. The page didn´t give us more data then their time of reign, so we only included these two colums.

Result: see appendix

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

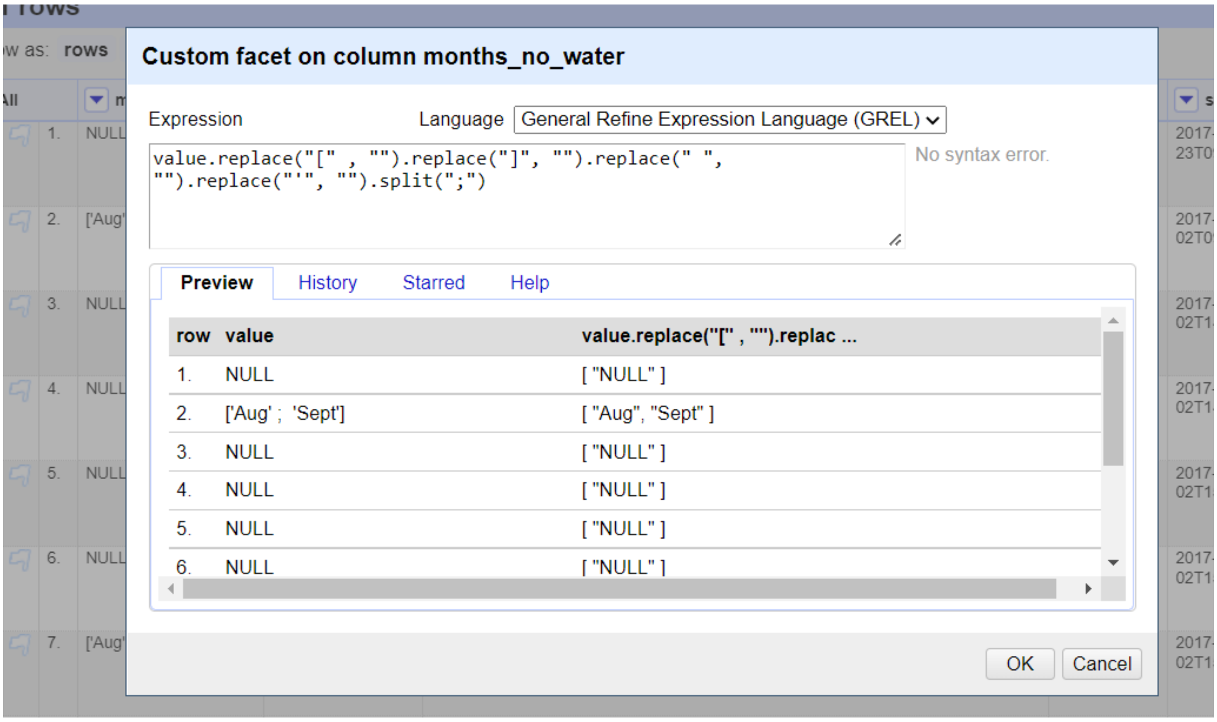
OpenRefine will not alter the raw data when using the sorting and filtering functions. It will only show a preview of your changes on the right site menu, so you will also have an overview of the original data. OpenRefine also includes an automated versioning, so it is always possible to go back to the previous version.

3. *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?"*

To figure out, which two months are reported the most water-deprived months according to the farmer households, I first have to find a column with the relevant data. There is a column called “Months_no_water”, which probably includes the information I need.

First I moved the column to the start of the data-set for better overview and inspected the data entries with a simple “text-facet”, which shows me a summary of all the data entries and how often they are counted in the data-set.

To fix that, I create a custom text facet where I type this function-call in: “value.replace("[" , "").replace("]", "").replace(" ", "").replace("'", "").split(";")”



Custom facet on column months_no_water

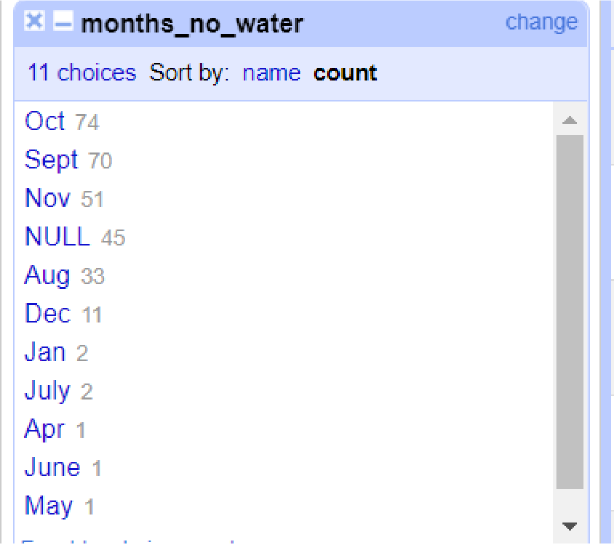
Expression   Language: General Refine Expression Language (GREL)

```
value.replace("[" , "").replace("]", "").replace(" ",
"").replace("'", "").split(";")
```

No syntax error.

row	value	value.replace("[" , "").replac ...
1.	NULL	["NULL"]
2.	['Aug' ; 'Sept']	["Aug", "Sept"]
3.	NULL	["NULL"]
4.	NULL	["NULL"]
5.	NULL	["NULL"]
6.	NULL	["NULL"]

This will now give me a preview of all the values by their own and how often they occur.



months_no_water   change

11 choices Sort by: name **count**

Oct	74
Sept	70
Nov	51
NULL	45
Aug	33
Dec	11
Jan	2
July	2
Apr	1
June	1
May	1

Where the answer of the two most water-deprived months is: October and September.

4. *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)

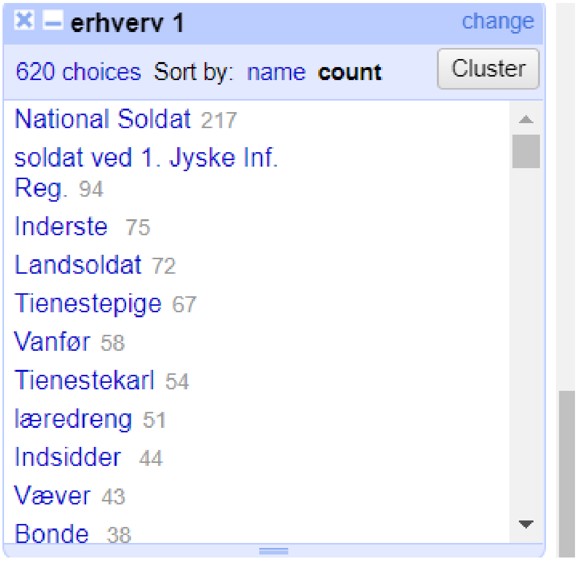
First I put the data-set into OpenRefine and inspected the columns and information.

To answer the question, I need the column “civilstand” and the column “erhverv”. I can already see, that in the column “erhverv” are multiple entries per row sometimes, separated by a comma or the Danish word “og”. So first of all, I would split the column into multiple columns, where I separate them by this regular expression “,|\bog”, which splits it at a comma or a the specific word “og”.

Now I 5 columns, called “erhverv 1-5”, where I just assume that the first occupation mentioned is their main occupation. After making a text facet on each of the new columns, it seems to be valid to assume this, as all the entries of the other columns would not have an effect on the “erhverv 1” column.

To now see, which occupation is the most frequent for unmarried men and women, I have to apply a text filter with the input “ugift”. Now it will only count in the entries with the people that have the status of unmarried.

Now I can make a normal text-facet on the “erhverv 1” column, and sort it by count. I then check the cluster, to see if there are any similar texts or entries, and there were a lot, which I then clustered accordingly. At last I will have a list of the 10 most frequent occupations among unmarried people in 1801 Aarhus:



erhverv 1

620 choices

National Soldat 217
soldat ved 1. Jyske Inf. Reg. 94
Inderste 75
Landsoldat 72
Tienestepige 67
Vanfør 58
Tienestekarl 54
læredreng 51
Indsidder 44
Væver 43
Bonde 38