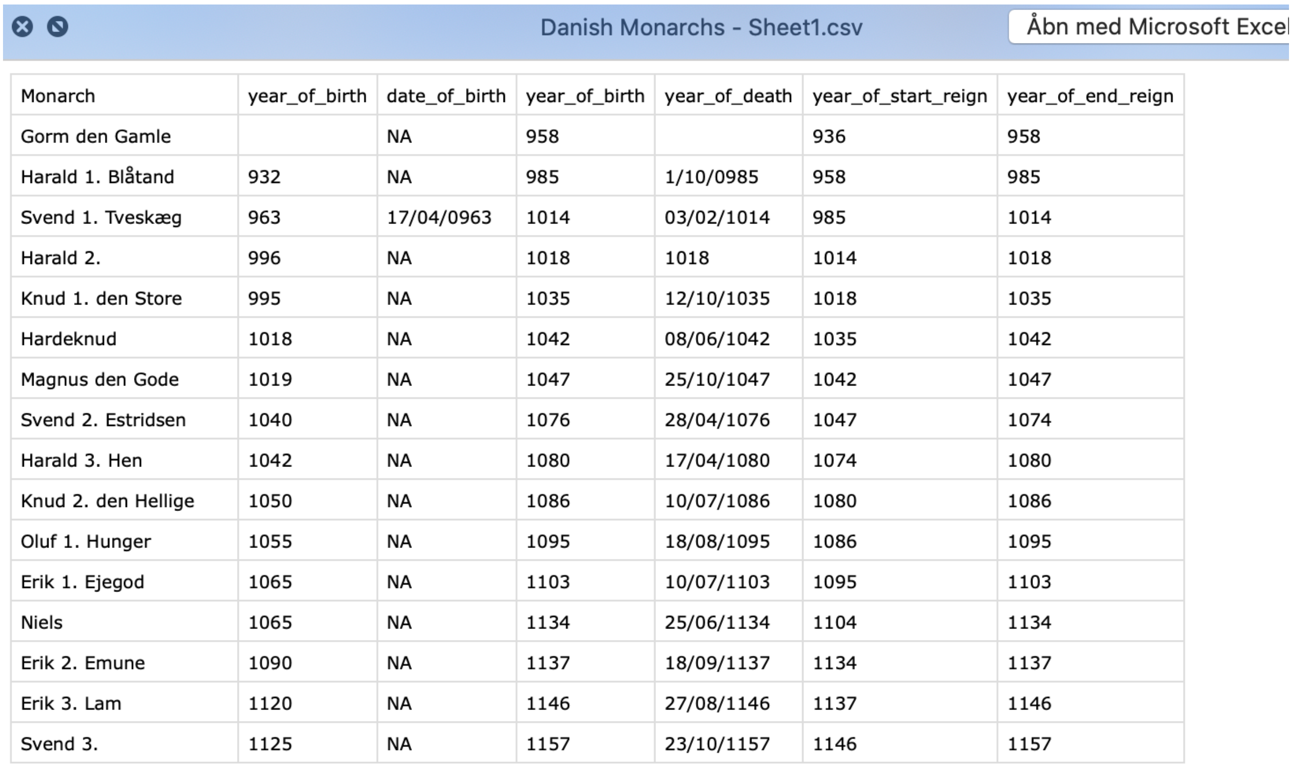
2:W35: Open Refine

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.

1.1 ANSWER



Monarch	year_of_birth	date_of_birth	year_of_birth	year_of_death	year_of_start_reign	year_of_end_reign
Gorm den Gamle		NA	958		936	958
Harald 1. Blåtand	932	NA	985	1/10/0985	958	985
Svend 1. Tveskæg	963	17/04/0963	1014	03/02/1014	985	1014
Harald 2.	996	NA	1018	1018	1014	1018
Knud 1. den Store	995	NA	1035	12/10/1035	1018	1035
Hardeknud	1018	NA	1042	08/06/1042	1035	1042
Magnus den Gode	1019	NA	1047	25/10/1047	1042	1047
Svend 2. Estridsen	1040	NA	1076	28/04/1076	1047	1074
Harald 3. Hen	1042	NA	1080	17/04/1080	1074	1080
Knud 2. den Hellige	1050	NA	1086	10/07/1086	1080	1086
Oluf 1. Hunger	1055	NA	1095	18/08/1095	1086	1095
Erik 1. Ejegod	1065	NA	1103	10/07/1103	1095	1103
Niels	1065	NA	1134	25/06/1134	1104	1134
Erik 2. Emune	1090	NA	1137	18/09/1137	1134	1137
Erik 3. Lam	1120	NA	1146	27/08/1146	1137	1146
Svend 3.	1125	NA	1157	23/10/1157	1146	1157

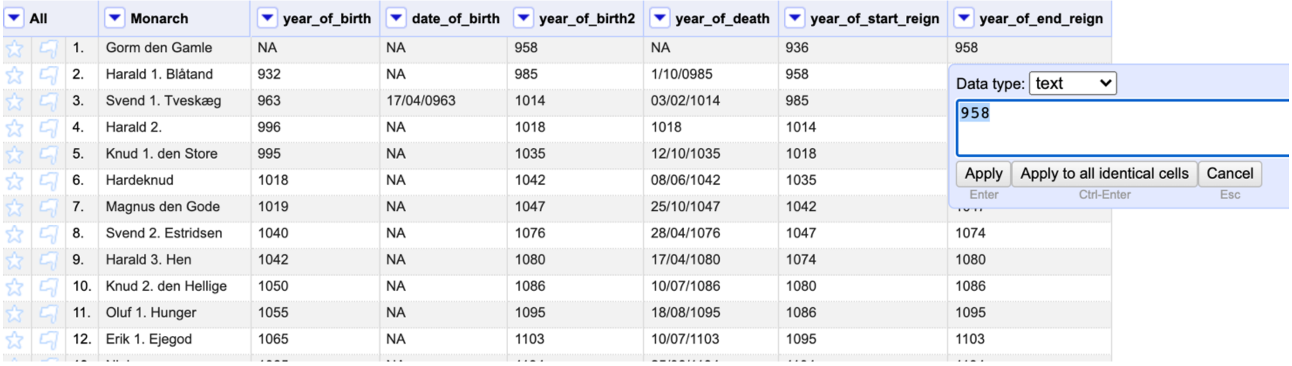
…… this is only a subset of the sheet

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

2.1 ANSWER

When opened in OpenRefine (2.1.1) all values seem to have been altered into text, which is not particularly easy convenient since no calculations (sum, count, etc.) can be done. For comparison all the values but the name of the monarchs and date of birth, which turns into characters, are integers when loaded into RStudio (2.1.2) – not the most convenient either, but easier (for me at least) to turn into numeric values.

2.1.1

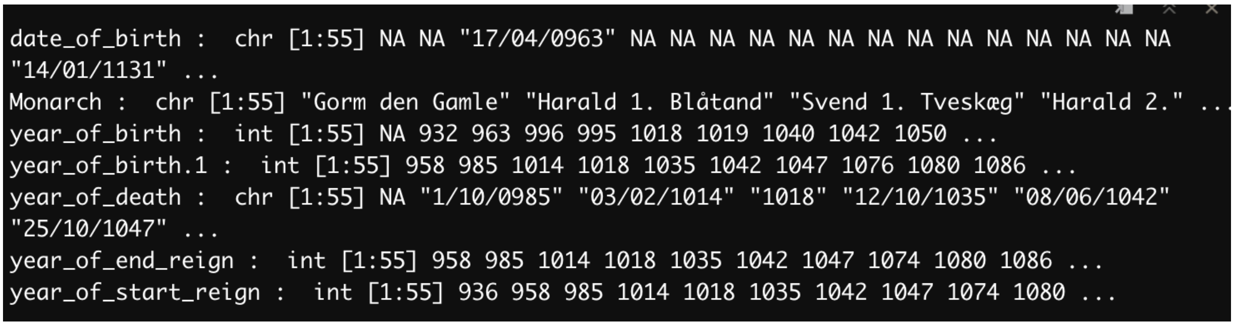


All	Monarch	year_of_birth	date_of_birth	year_of_birth2	year_of_death	year_of_start_reign	year_of_end_reign
1.	Gorm den Gamle	NA	NA	958	NA	936	958
2.	Harald 1. Blåtand	932	NA	985	1/10/0985	958	
3.	Svend 1. Tveskæg	963	17/04/0963	1014	03/02/1014	985	
4.	Harald 2.	996	NA	1018	1018	1014	
5.	Knud 1. den Store	995	NA	1035	12/10/1035	1018	
6.	Hardeknud	1018	NA	1042	08/06/1042	1035	
7.	Magnus den Gode	1019	NA	1047	25/10/1047	1042	
8.	Svend 2. Estridsen	1040	NA	1076	28/04/1076	1047	1074
9.	Harald 3. Hen	1042	NA	1080	17/04/1080	1074	1080
10.	Knud 2. den Hellige	1050	NA	1086	10/07/1086	1080	1086
11.	Oluf 1. Hunger	1055	NA	1095	18/08/1095	1086	1095
12.	Erik 1. Ejegod	1065	NA	1103	10/07/1103	1095	1103

Data type: text

958

2.1.2



```
date_of_birth :  chr [1:55] NA NA "17/04/0963" NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
"14/01/1131" ...
Monarch :  chr [1:55] "Gorm den Gamle" "Harald 1. Blåtand" "Svend 1. Tveskæg" "Harald 2." ...
year_of_birth :  int [1:55] NA 932 963 996 995 1018 1019 1040 1042 1050 ...
year_of_birth.1 :  int [1:55] 958 985 1014 1018 1035 1042 1047 1076 1080 1086 ...
year_of_death :  chr [1:55] NA "1/10/0985" "03/02/1014" "1018" "12/10/1035" "08/06/1042"
"25/10/1047" ...
year_of_end_reign :  int [1:55] 958 985 1014 1018 1035 1042 1047 1074 1080 1086 ...
year_of_start_reign :  int [1:55] 936 958 985 1014 1018 1035 1042 1047 1074 1080 ...
```

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?"

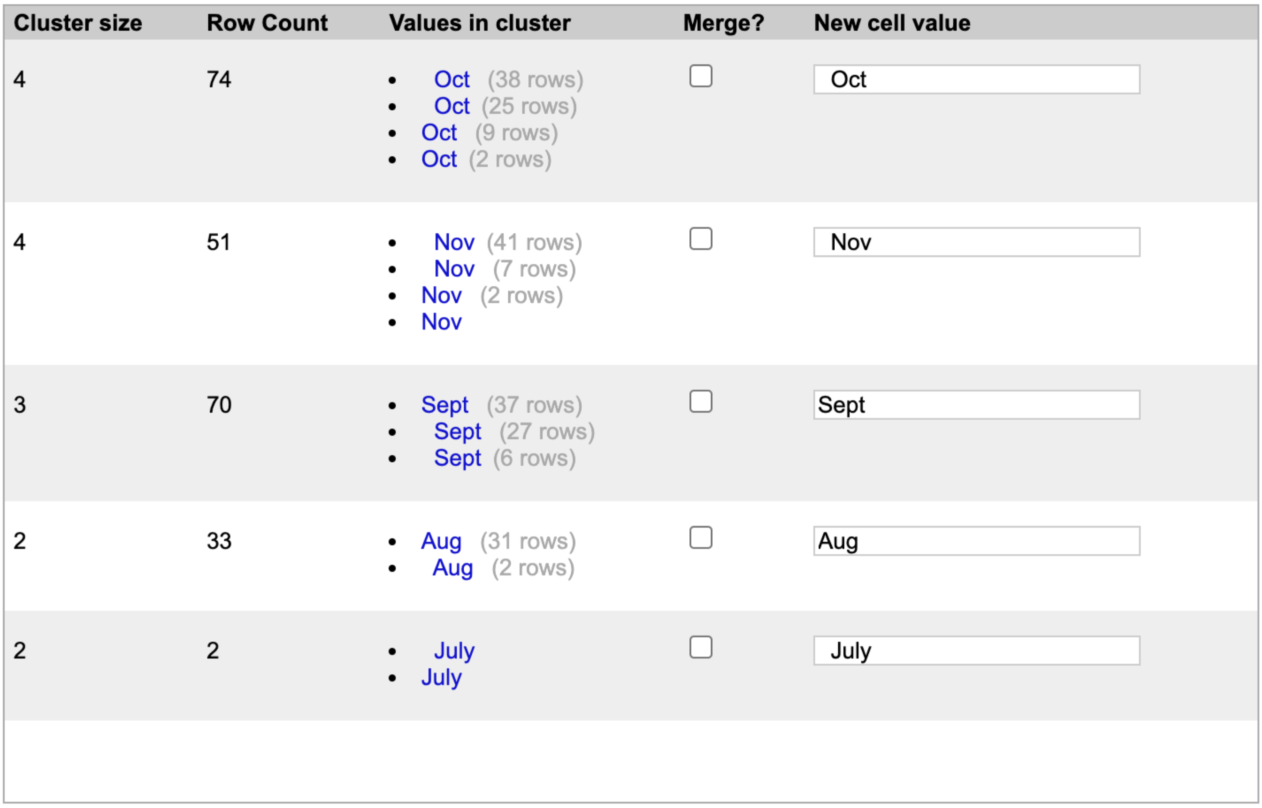
3.1 ANSWER

To answer this the following has been done:
I look at the column *months_no_water* and count which months are the most reported. These are originally reported like this [‘month’ ; ‘month’], thus the data must first be cleaned. To do so I first press the cell then:

- Edit cell
- Split multivalued cell
    - ○ Enter **;** to split the cell by this
- Then I press the cell again, followed by edit cell and transform
    - ○ Custom text transformer
    - ○ Enter *value.replace("[","")* and *value.replace("]","")* and value.replace("’","")

Then I pressed Facet – text facet and cluster, to count which months are the most reported. As 3.1.1. shows **October (74)** and **September (70)** are the most reported water-deprived months.

3.1.1



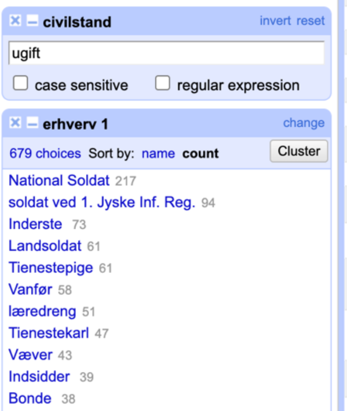
Cluster size	Row Count	Values in cluster	Merge?	New cell value
4	74	• Oct (38 rows) • Oct (25 rows) • Oct (9 rows) • Oct (2 rows)	□	Oct
4	51	• Nov (41 rows) • Nov (7 rows) • Nov (2 rows) • Nov	□	Nov
3	70	• Sept (37 rows) • Sept (27 rows) • Sept (6 rows)	□	Sept
2	33	• Aug (31 rows) • Aug (2 rows)	□	Aug
2	2	• July • July	□	July

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)

4.1 ANSWER

To answer this the following has been done:
I look at the column *erhverv* which are messy reported, thus the data must first be cleaned. To do so I first press the cell then:

- Edit cell
- Split multivalued cell
  - Enter *og* to split the words by this
- facet for the *civilstand* and change it to *ugift* see 4.1.1 to ensure only the occupation of unmarried people were considered
- facet for *erhverv*
  - cluster across similar titles



ugift

□ case sensitive □ regular expression

679 choices Sort by: name count

National Soldat	217
soldat ved 1. Jyske Inf. Reg.	94
Inderste	73
Landsoldat	61
Tienestepige	61
Vanfør	58
læredreng	51
Tienestekarl	47
Væver	43
Indsidder	39
Bonde	38

4.1.1

According to this the five most frequent occupations among unmarried people in 1801 are National Soldat (217), Soldat wed 1. Jyske Inf. Reg (94), Inderste (73) and Landsoldat (61) and Tjenestepige (61), which suggest that a great majority of the people were some kind of soldier. One should however note that a lot of observations are missing, and this could indicate that mainly male dominated occupations are reported.