Homework 2 - Open Refine

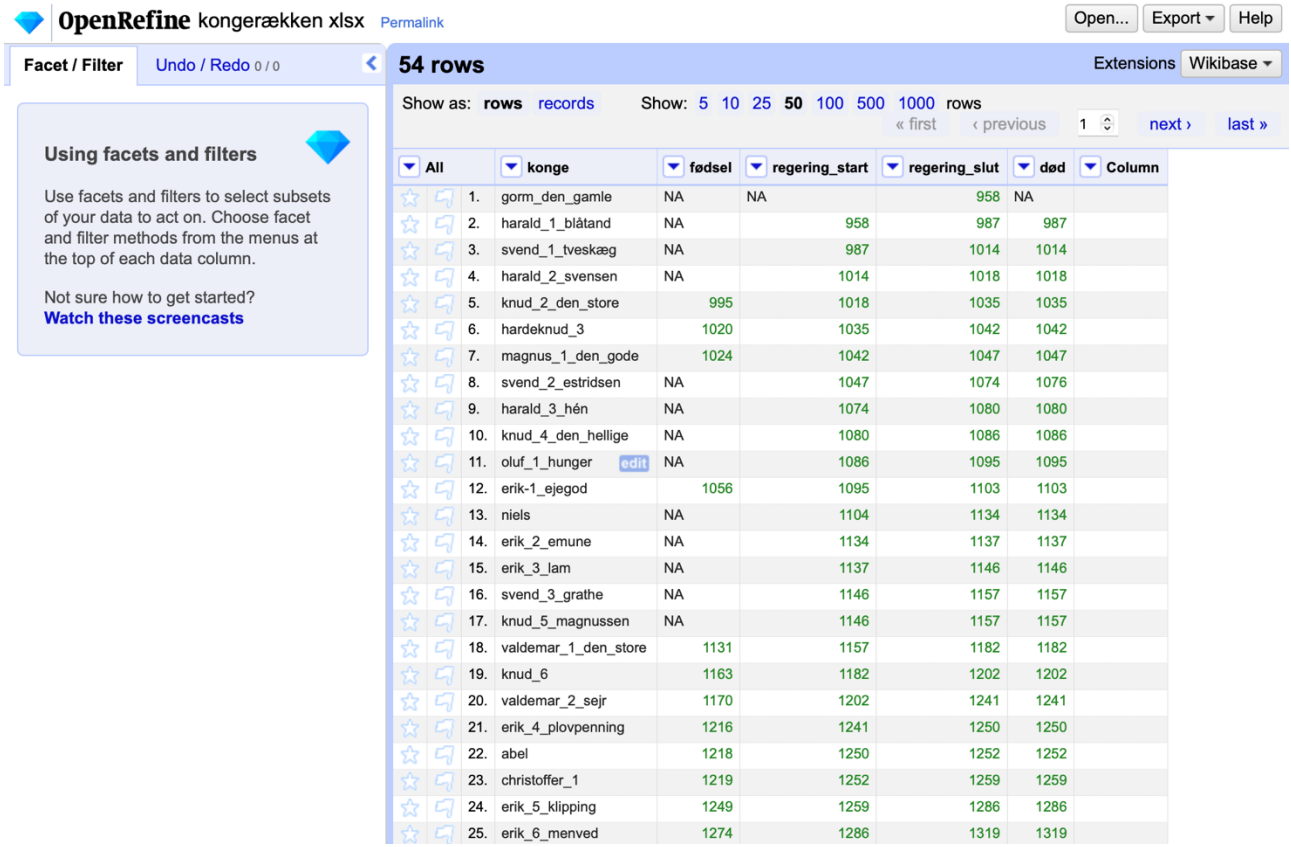
1. **Create a *tidy* spreadsheet/table listing the names of Danish monarchs with their birth- and death-date and start and end of their reign. They should be sortable by year of birth. Suitable source website is for example here, but you can also use another source, provided you reference it. (Collaboration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

- I made a ”tidy dataset” as formulated by Hadley Wickham, in excel over the Danish kings and queens row. I did it based on the information found on the following page: https://danmarkshistorien.dk/vis/materiale/kongeraekken
    - The kings are sorted, as far as it has been possible (due to lack of data) by their birth date, in accordance with the information available at the link above.
    - Because Christopher II have been king 2 different periods, he can be found twice at the list.
    - For some kings, the exact year of birth is not known, but then i based the order of the kings on the list by the estimated start of their reign.
- Measures used to create a good spreadsheet.
    - I followed the “tidy data” format.
    - Empty spots (missing data) are filled with NA, to avoid messing up the data with numbers
    - Instead of spaces, i have used underscore in the names of the kings and in the headlines as well.
    - Both the number of the kings and their nickname are provided (e.g. Svend_1_tveskæg)

konge	fødsel	regering_start	regering_slut	død
gorm_den_gamle	NA	NA	958	NA
harald_1_blåtand	NA	958	987	987
svend_1_tveskæg	NA	987	1014	1014
harald_2_svensen	NA	1014	1018	1018
knud_2_den_store	995	1018	1035	1035
hardeknud_3	1020	1035	1042	1042
magnus_1_den_gode	1024	1042	1047	1047
svend_2_estridsen	NA	1047	1074	1076
harald_3_hén	NA	1074	1080	1080
knud_4_den_hellige	NA	1080	1086	1086
oluf_1_hunger	NA	1086	1095	1095
erik-1_ejegod	1056	1095	1103	1103
niels	NA	1104	1134	1134
erik_2_emune	NA	1134	1137	1137
erik_3_lam	NA	1137	1146	1146
svend_3_grathe	NA	1146	1157	1157

knud_5_magnussen	NA	1146	1157	1157
valdemar_1_den_store	1131	1157	1182	1182
knud_6	1163	1182	1202	1202
valdemar_2_sejr	1170	1202	1241	1241
erik_4_plovpenning	1216	1241	1250	1250
abel	1218	1250	1252	1252
christoffer_1	1219	1252	1259	1259
erik_5_klipping	1249	1259	1286	1286
erik_6_menved	1274	1286	1319	1319
christoffer_2	1276	1320	1326	1326
christoffer_2	1276	1330	1332	1326
valdemar_3_eriksen	1314	1326	1330	1330
valdemar_4_atterdag	1320	1340	1375	1375
margrete_1	1353	1387	1412	1412
oluf_2	1370	1376	1387	1387
erik_7_af_pommeren	1382	1412	1439	1459
christoffer_3_af_bayern	1416	1440	1448	1448
christian_1	1426	1448	1481	1481
hans	1455	1481	1513	1513
frederik_1	1471	1523	1533	1533
christian_2	1481	1513	1523	1559
christian_3	1503	1534	1559	1559
frederik_2	1534	1559	1588	1588
christian_4	1577	1588	1648	1648
frederik_3	1609	1648	1670	1670
christian_5	1646	1670	1699	1699
frederik_4	1671	1699	1730	1730
christian_6	1699	1730	1746	1746
frederik_5	1723	1746	1766	1766
christian_7	1749	1766	1808	1808
frederik_6	1768	1808	1839	1839
christian_8	1786	1839	1848	1848
frederik_7	1808	1848	1863	1863
christian_9	1818	1863	1906	1906
frederik_8	1843	1906	1912	1912
christian_10	1870	1912	1947	1947
frederik_9	1899	1947	1972	1972
margrethe_2	1940	1972	NA	NA

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

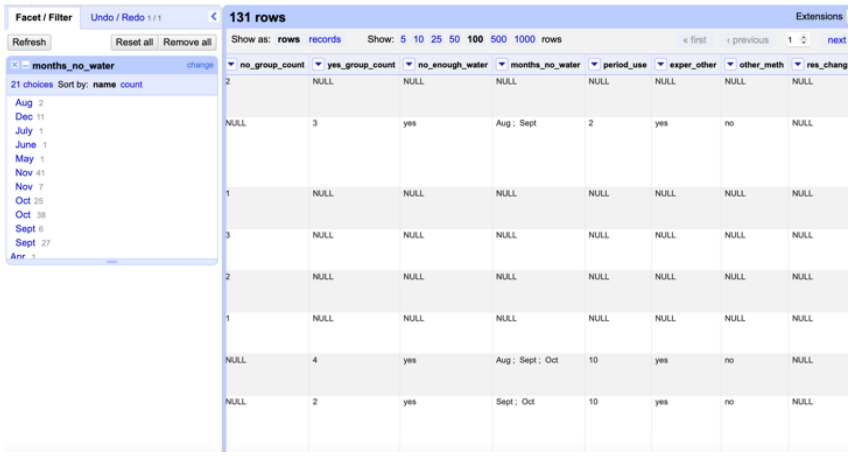


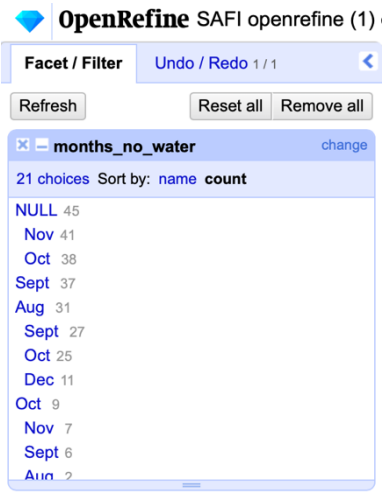
All		konge	fødsel	regering_start	regering_slut	død	Column
	1.	gorm_den_gamle	NA	NA	958	NA	
	2.	harald_1_blåtand	NA	958	987	987	
	3.	svend_1_tveskæg	NA	987	1014	1014	
	4.	harald_2_svensen	NA	1014	1018	1018	
	5.	knud_2_den_store	995	1018	1035	1035	
	6.	hardeknud_3	1020	1035	1042	1042	
	7.	magnus_1_den_gode	1024	1042	1047	1047	
	8.	svend_2_estridsen	NA	1047	1074	1076	
	9.	harald_3_hén	NA	1074	1080	1080	
	10.	knud_4_den_hellige	NA	1080	1086	1086	
	11.	oluf_1_hunger edit	NA	1086	1095	1095	
	12.	erik-1_ejegod	1056	1095	1103	1103	
	13.	niels	NA	1104	1134	1134	
	14.	erik_2_emune	NA	1134	1137	1137	
	15.	erik_3_lam	NA	1137	1146	1146	
	16.	svend_3_grathe	NA	1146	1157	1157	
	17.	knud_5_magnussen	NA	1146	1157	1157	
	18.	valdemar_1_den_store	1131	1157	1182	1182	
	19.	knud_6	1163	1182	1202	1202	
	20.	valdemar_2_sejr	1170	1202	1241	1241	
	21.	erik_4_plovpenning	1216	1241	1250	1250	
	22.	abel	1218	1250	1252	1252	
	23.	christoffer_1	1219	1252	1259	1259	
	24.	erik_5_klipping	1249	1259	1286	1286	
	25.	erik_6_menved	1274	1286	1319	1319	

No, it seems to work just fine in open refine, it only changes the years i have written to date format.

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?"**

I took the column ”month_no_water”, and used the regular expression: value.replace("[","").replace("]","").replace("'","") to separate the data in each cell, so that they count individually.
I went to text facet, then pressed the box ”change” and used the regular expression: value.split”;”
I ended up with the following spread sheet.





Now i could just used ”count” and ignored the ”NULL” values.
In conclusion it therefore seems that it is November and October that have been mentioned as the driest months.

4. **OPTIONAL 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. As an inspiration, check out this chapter Making a living outside marriage from the Swedish Gender and Work project of Maria Agren)**

I tried:
I uploaded the link in Openrefine
I then selected the following rows and created a 'Text Facet':

- Farmstand
- Erhverv (Profession)
- Civilstand (Marital Status)
- Køn (Gender)

I then looked further into the unmarried women, which includes ”separated” ”widows” and just ”unmarried”. Therefore, i deselected blank spots and data from the married women’s. And then i started using the cluster function and created a more uniform data sample.
Unfortunately, I haven't completed it yet, but it appears that unmarried women either have domestic occupations or work with handcrafts such as weaving, sewing, or spinning (I grouped these into the same category).