Gruppe: Kristoffer Segerstrøm, Emil Gert Hansen, Lukas Benner

Week 8 assignment

1. **What regular expressions do you use to extract all the dates in this blurb: http://bit.ly/regexexercise2 and to put them into the following format YYYY-MM-DD ?**

When we want to convert dates to the same form, we first need to mark all the dates.

We can use these commands to mark the dates in the text:

\d+.\d+.\d+

or

\d{1,2}.\d{1,2}.\s?\d+

When we have marked the dates, we put a bracket around every group:

(\d{1,2}).(\d{1,2}).\s?(\d+)

Then go to function in the right colon and find substitution.

In the textbox that opens with substitution write:

$2-$1-$3

This follows the order in which the date is shown. Here the date is formed as number 2, 1 is the month and 3 represents the year.. This can be written in any wanted order.

The task in regex.

https://regex101.com/r/83vEpt/1

2. **Write a regular expression to convert the stopwordlist (list of most frequent Danish words) from Voyant in http://bit.ly/regexexercise3 into a neat stopword list for R (which comprises "words" separated by commas, such as http://bit.ly/regexexercise4). Then take the stopwordlist from R http://bit.ly/regexexercise4 and convert it into a Voyant list (words on separate line without interpunction)**

Stoplist from Voyant to R

Write ([a-z0-9æøå]+) in the textbox followed by \n:

([a-z0-9æøå]+)\n

The \n makes the text go from colon to text.

Go to function and substitution. Write the following

"$1"

Then you have a stoplist for R

The task in Regex:

https://regex101.com/r/WQtOTA/1

Stoplist from R to Voyant

Write the following in the textbox:

\"([a-z0-9æøåüé.]+)(.,)

\" and (.,) deletes the characters.

Go to function and substitution. Write the following

$1\n

The \n makes the text go from text to colon.

Then you have a stoplist for Voyant.

The task in Regex:

https://regex101.com/r/zTCekK/1

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

   No, OpenRefine does not alter the raw data when you sort or filter it. Sorting and filtering in OpenRefine are non-destructive operations that only change the way data is displayed but do not modify the underlying dataset.

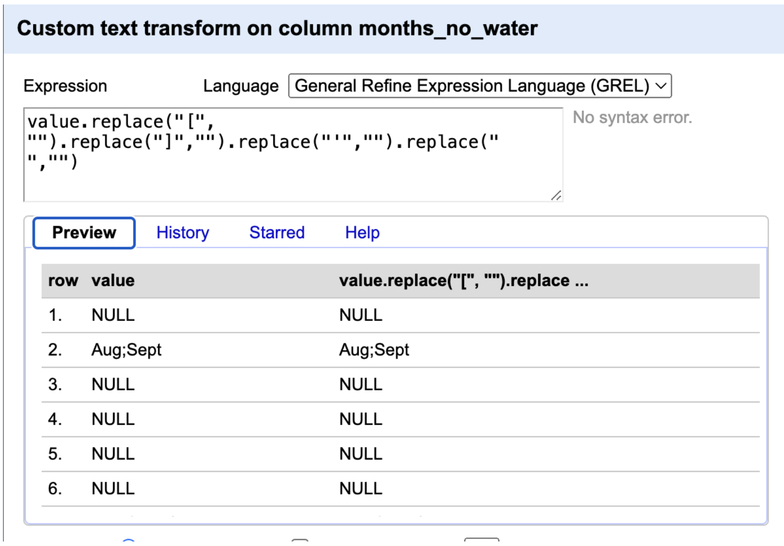
Gruppe: Kristoffer Segerstrøm, Emil Gert Hansen, Lukas Benner

4. **Fix the interviews dataset in OpenRefine enough to answer this question: "Which two months are reported as the most water-deprived/dryest by the interviewed farmer households?"**

The two most water-deprived months are October and September.

This conclusion is made by using the command in OpenRefine, where we choose the colon "months_lack_water".

We choose "facet" and then "transformation"



Custom text transform on column months_no_water

Expression Language: General Refine Expression Language (GREL)

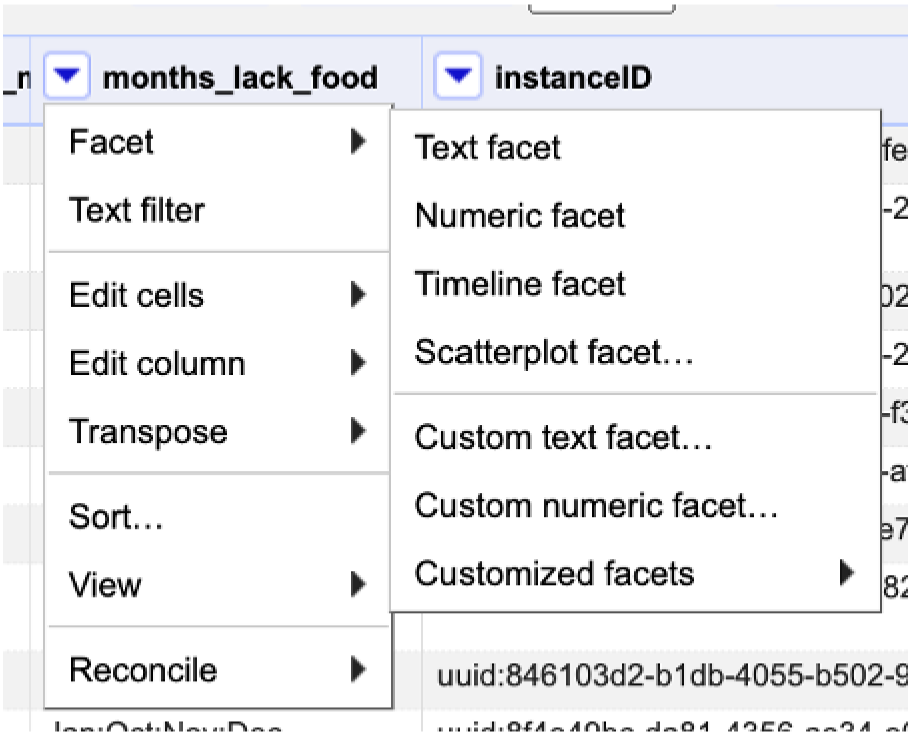
```
value.replace("[",
"").replace("]","").replace("'","").replace("
","")
```

No syntax error.

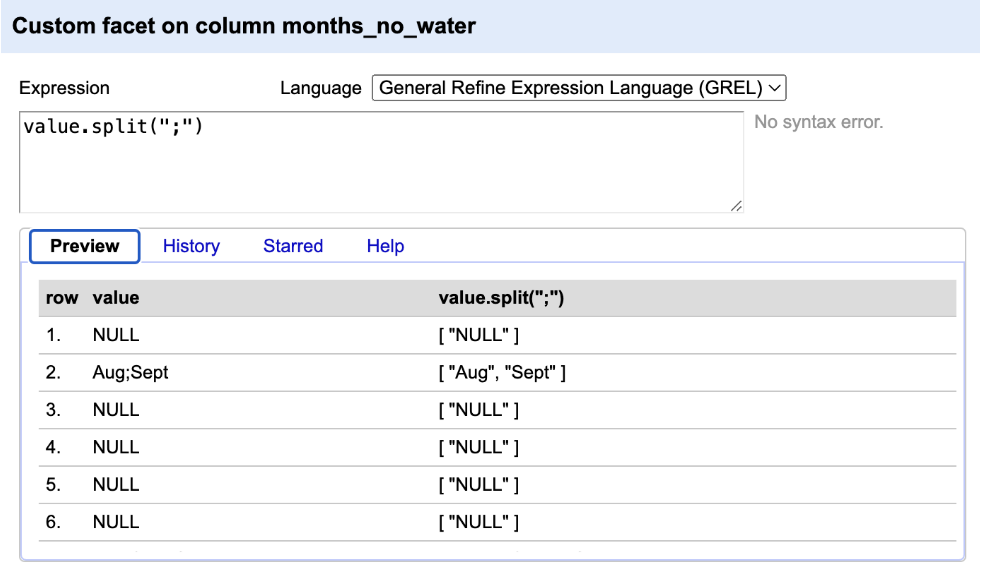
row	value	value.replace("[", "").replace ...
1.	NULL	NULL
2.	Aug;Sept	Aug;Sept
3.	NULL	NULL
4.	NULL	NULL
5.	NULL	NULL
6.	NULL	NULL

Then we choose "facet" and then "custom text facet"

Gruppe: Kristoffer Segerstrøm, Emil Gert Hansen, Lukas Benner



In “Custom text facet” we write “value.split(“;”)



Custom facet on column months_no_water

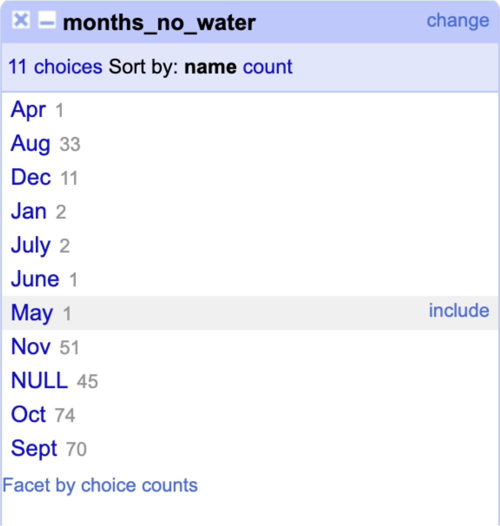
Expression Language General Refine Expression Language (GREL)

```
value.split(";")
```

No syntax error.

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

We now get a facet, where we can see a list of the months ranged and how many that experienced lack of water these months.



months_no_water

11 choices Sort by: **name** count

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

From the list we can see that in October there were 74 that lacked food and 70 in September.

5. **Real-Data Challenge: What are the 10 most frequent occupations "erhverv" among unmarried men or women of 20-30 years in 1801 Aarhus census dataset? (hint: first select either men or women to shrink the dataset to a manageable size, then filter by age, and then use merging to cut the erhverv variation ruthlessly.)**

We chose to focus on the women.

Women:

First, we sort by gender:

Facet → text facet

Then we sort by age:

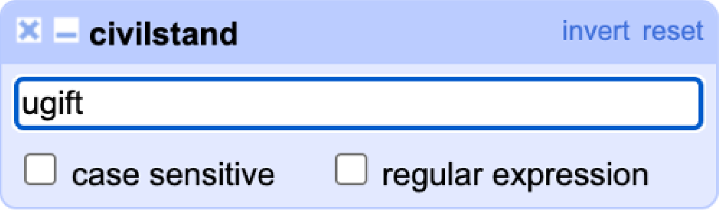
Edit cells → Common transforms → To number

Facet → Numeric facet

Then age can then be adjusted to 20-30

To separate the married and unmarried from each other we create a text filter under “civilstand” and write “ugift” in the facet, which appears.

Gruppe: Kristoffer Segerstrøm, Emil Gert Hansen, Lukas Benner



civilstand

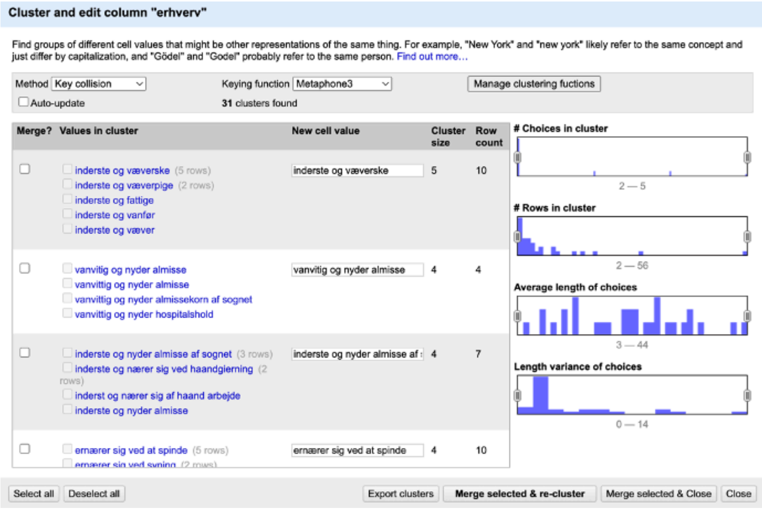
invert reset

ugift

□ case sensitive □ regular expression

Then we make a facet of “erhverv” and cluster the words that can be clustered.

First with the methods of “Key collision” and keying function “Metaphone3” where we re-merge and cluster



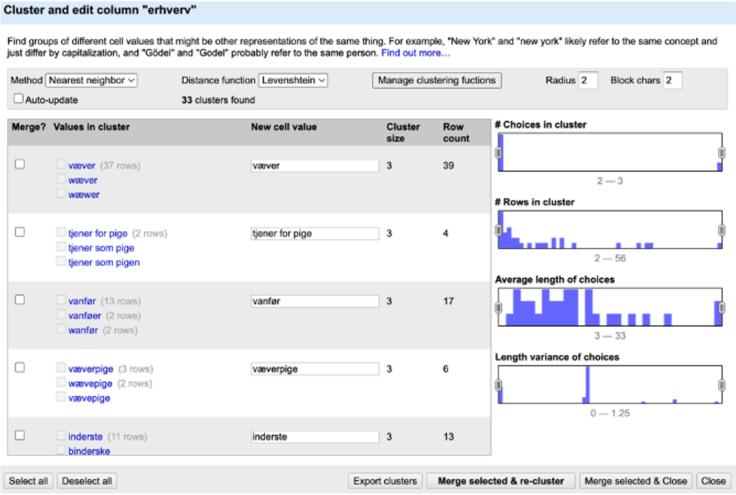
Cluster and edit column "erhverv"

Find groups of different cell values that might be other representations of the same thing. For example, "New York" and "new york" likely refer to the same concept and just differ by capitalization, and "Gödel" and "Godel" probably refer to the same person. Find out more…

Method: Key collision   Keying function: Metaphone3   Manage clustering fuctions
□ Auto-update   31 clusters found

Merge?	Values in cluster	New cell value	Cluster size	Row count
□	□ inderste og væverske (5 rows) □ inderste og væverpige (2 rows) □ inderste og fattige □ inderste og vanfør □ inderste og væver	inderste og væverske	5	10
□	□ vanvitig og nyder almisse □ vanvittig og nyder almisse □ vanvittig og nyder almissekorn af sognet □ vanvittig og nyder hospitalshold	vanvitig og nyder almisse	4	4
□	□ inderste og nyder almisse af sognet (3 rows) □ inderste og nærer sig ved haandgierning (2 rows) □ inderst og nærer sig af haand arbejde □ inderste og nyder almisse	inderste og nyder almisse af :	4	7
□	□ ernærer sig ved at spinde (5 rows) □ ernærer sig ved spning (2 rows)	ernærer sig ved at spinde	4	10

Afterwards we re-merge and cluster with “Nearest neighbor” and a radius within 2 and block chars 2



Cluster and edit column "erhverv"

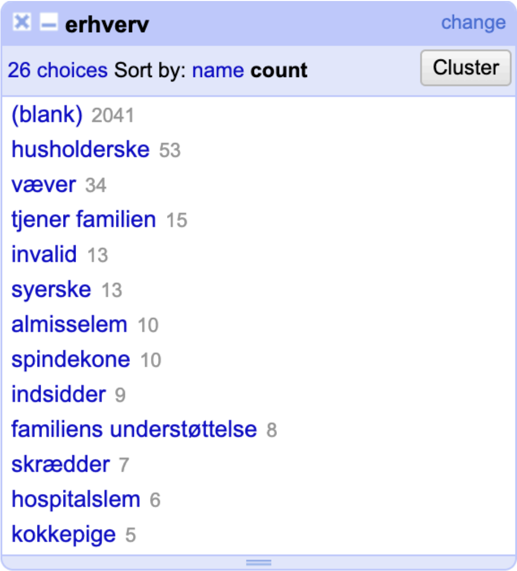
Find groups of different cell values that might be other representations of the same thing. For example, "New York" and "new york" likely refer to the same concept and just differ by capitalization, and "Gödel" and "Godel" probably refer to the same person. Find out more…

Method: Nearest neighbor   Distance function: Levenshtein   Manage clustering fuctions   Radius: 2   Block chars: 2
□ Auto-update   33 clusters found

Merge?	Values in cluster	New cell value	Cluster size	Row count
□	□ væver (37 rows) □ wæver □ wæwer	væver	3	39
□	□ tjener for pige (2 rows) □ tjener som pige □ tjener som pigen	tjener for pige	3	4
□	□ vanfør (13 rows) □ vanføer (2 rows) □ wanfør (2 rows)	vanfør	3	17
□	□ væverpige (3 rows) □ wævepige (2 rows) □ vævepige	væverpige	3	6
□	□ inderste (11 rows) □ binderske	inderste	3	13

Gruppe: Kristoffer Segerstrøm, Emil Gert Hansen, Lukas Benner

The rest we edit manually and are sorted roughly into bigger groups seen below.

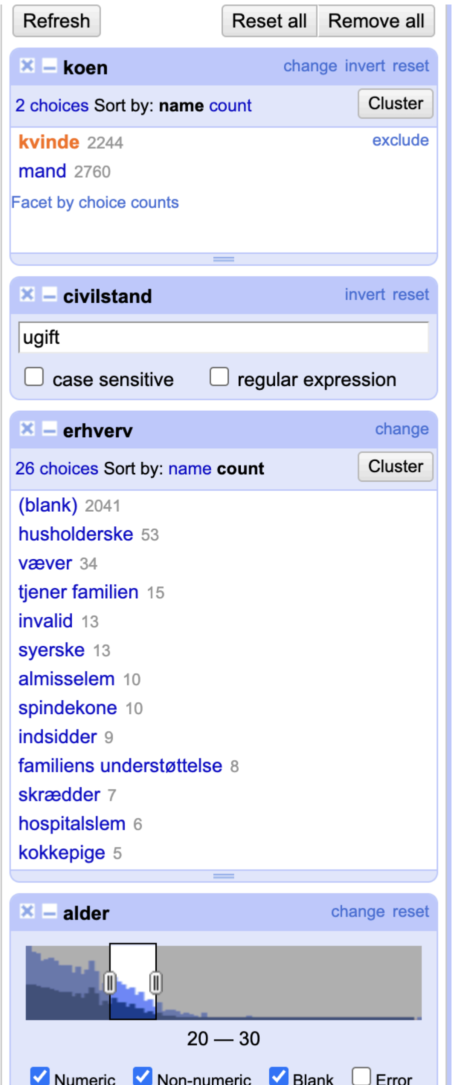


erhverv change

26 choices Sort by: name **count** Cluster

(blank) 2041
husholderske 53
væver 34
tjener familien 15
invalid 13
syerske 13
almisselem 10
spindekone 10
indsidder 9
familiens understøttelse 8
skrædder 7
hospitalslem 6
kokkepige 5

This shows the top 10 jobs for women, where maid is the main occupation.



Refresh Reset all Remove all

koen change invert reset

2 choices Sort by: **name** count Cluster

kvinde 2244 exclude
mand 2760

Facet by choice counts

civilstand invert reset

ugift

□ case sensitive □ regular expression

erhverv change

26 choices Sort by: name **count** Cluster

(blank) 2041
husholderske 53
væver 34
tjener familien 15
invalid 13
syerske 13
almisselem 10
spindekone 10
indsidder 9
familiens understøttelse 8
skrædder 7
hospitalslem 6
kokkepige 5

alder change reset

20 — 30

☑ Numeric ☑ Non-numeric ☑ Blank □ Error

This is the overview of the facets.