Homework Week 2

Anders Hjulmand

08/09/2021

Github: https://github.com/ah140797/CulturalDataScience

---

1 Does OpenRefine alter the raw data during sorting and filtering?

The short answer is no. When sorting, the order of the rows changes by a coloumn variable. Imagine an experiment where you have one participant per row and several variables in the coloumns. You can sort this dataframe by age so that participants with the lowest (or highest) age appears at the first rows. This doesn't change the raw data, it merely reorders it. The same logic applies for filtering, where rows are selected based on selected boolean criterion.

2 Which two months are reported as the most water-deprived/driest by the interviewed farmer households

October and Semtember were reported to be the most water-deprived months by the farmer households. First i used the `custom text facet` with the expression `value.replace` to remove the the following symbols `[ ] '`. Then i split the remaining words with `split multi-varied cells` using `;` as the separator. Finally i clustered the words.

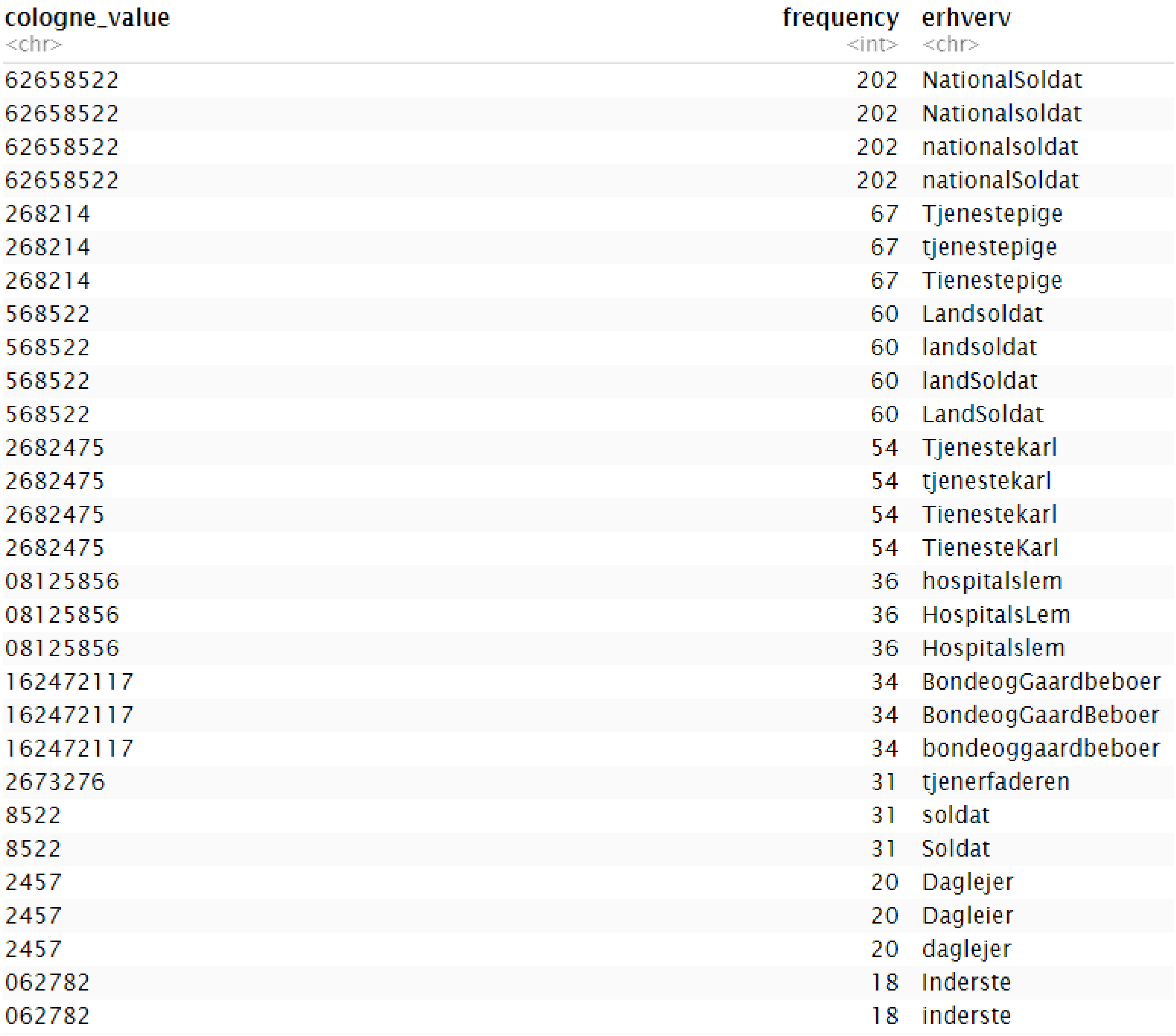
3 What were the 10 most frequent occupations (erhverv) among unmarried men and women in 1801 Aarhus?

Figure 1 shows a list of the 10 most frequent occupations among unmarried men and women in 1801 Aarhus. I used `R` to make this exercise. See Github for the code. First i used the function `filter` from the `tidyverse package` to only include rows with ugifte subjects. I then removed rows that had missing values in the coloumn "evhverv" using the function `drop_na` and removed spaces in the coloumn "erhverv" using the function `str_replace_all`. I then added a new coloumn named `cologne_value` which contains a sequence of digits representing the phonetic code using the package `phonics`.

```
df_ugift <- df %>%
  #filtering ugifte
  filter(civilstand == "ugift") %>%
  #removing na's
  drop_na(erhverv) %>%
  mutate(
    erhverv = as.character(erhverv),
    #removing spaces
    erhverv = str_replace_all(erhverv, "␣", ""),
    #adding a coloumn with cologne-phonetic values for "erhverv"
    cologne_value = cologne(erhverv),
  )
```

I continued by using the function `group_by` combined with `sumamrize` to count the frequency of each unique cologne value. I also added the erhverv to make it more readable. It seems like the `phonics package` does actually work well. See figure 1.

```
df_ugift %>%
  drop_na(cologne_value) %>%
  group_by(cologne_value) %>%
  summarize(
            frequency = length(cologne_value),
            erhverv = unique(erhverv)) %>%
  arrange(desc(frequency))
```



cologne_value	frequency	erhverv
<chr>	<int>	<chr>
62658522	202	NationalSoldat
62658522	202	Nationalsoldat
62658522	202	nationalsoldat
62658522	202	nationalSoldat
268214	67	Tjenestepige
268214	67	tjenestepige
268214	67	Tienestepige
568522	60	Landsoldat
568522	60	landsoldat
568522	60	landSoldat
568522	60	LandSoldat
2682475	54	Tjenestekarl
2682475	54	tjenestekarl
2682475	54	Tienestekarl
2682475	54	TienesteKarl
08125856	36	hospitalslem
08125856	36	HospitalsLem
08125856	36	Hospitalslem
162472117	34	BondeogGaardbeboer
162472117	34	BondeogGaardBeboer
162472117	34	bondeoggaardbeboer
2673276	31	tjenerfaderen
8522	31	soldat
8522	31	Soldat
2457	20	Daglejer
2457	20	Dagleier
2457	20	daglejer
062782	18	Inderste
062782	18	inderste

Figure 1: The 10 most frequent occupations and their frequency