***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***

Et billede, der indeholder bord

Automatisk genereret beskrivelse

…… this is only a subset of the sheet

1. 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***

Et billede, der indeholder bord

Automatisk genereret beskrivelse

***2.1.2***

Et billede, der indeholder tekst

Automatisk genereret beskrivelse

1. Fix the [interviews dataset](https://ndownloader.figshare.com/files/11502815) 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**

Et billede, der indeholder bord

Automatisk genereret beskrivelse

1. Real-Data-Challenge: What are the 10 most frequent occupations (erhverv) among unmarried men and women in [1801 Aarhus](https://raw.githubusercontent.com/aarhusstadsarkiv/datasets/master/censuses/1801/census-1801-normalized.csv)? (hint: some expert judgement interpretation is necessary, look at the [HISCO classification](https://github.com/cedarfoundation/hisco) "Historical International Standard of Classification of Occupations" on [Dataverse](https://datasets.iisg.amsterdam/dataset.xhtml?persistentId=hdl:10622/88ZXD8) 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

**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.