2:W35: Open Refine

1. *See the attached spreadsheet document*
2. OpenRefine alters the raw data in the sense, that it adds a time stamp (time of the day) to some of the data fields. This can be explained by what was covered in the first lecture, that Excel is not great at handling dates when the year is too old (date before 1/1/1900), so it only adds a timepoint to the dates, that are after 1/1/1900.
   1. The above happens in the beginning, when the data is loaded into OpenRefine

If you only save a CSV-file after filtering/transforming the raw data, then the raw data is not exported along with the csv file, so be aware of this. The raw data and the different sorting/transformation steps can all be saved, if the data is saved as a tar.gz file

1. The two months that are reported as the most water-deprived/driest by the interviewed farmer households are **September** and **October.**
   1. The answer was found by clicking on the drop-down menu of the ‘monts\_no\_water’ column, then going to ‘Edit cells’ and then choosing ‘Split multi-valued cells…’. Consequently, the separator “;” was specified. Afterwards, in the overview box of ‘monts\_no\_water’ on the left side, the cells were clustered accordingly. Finally, the two most frequent months were found by sorting by count in the overview.
2. The 10 most frequent occupations among unmarried men and women in 1801 Aarhus are the following:
3. Nationalsoldat
4. Soldat ved 1. Jyske Inf. Reg.
5. Væver
6. Almisselem
7. Landsoldat
8. Tjenestekarl/tjenestepige
9. Spinder
10. Skrædder
11. Lærerdreng
12. Bonde