**Learning Journal weeks 6-8**

**Ellen Kirkpatrick**

**Week 6**

Open Refine data carpentry lessons

Working with openrefine

1. start openrefine. double click on .exe. file. Got taken to a blank black screen with lots of output getting written. It took some time but then openrefine came up in internet explorer.
2. Clicked create project, get data from, this computer. Selected SAFI messy file (had downloaded from data carpentry website). Uploaded data. Clicked create project.
3. Use faceting to look for potential errors. Scroll to village column, click down arrow and chose facet, text facet. Left panel a box appeared with all unique values in village column.
4. Sorted by name and count. data is problematic as is not consistent and some have spelling errors. Some have text names, others have number names.
5. Hovered mouse over names in facet list. Edit function becomes available.
6. Edited the errors so entries are all consistent. The name 49 cannot be made consistent with other data.
7. Applied text facet to interview date column. There are 19 options available. Column is formatted as text.
8. Produce timeline display for interview\_date. Select edit cells, common transfers, to date. Column converted to dates. Most dates were collected in November 2016.
9. Clustering - in village facet in left panel, click cluster. In pop up, selected key collision method and metaphone3 keying function. 2 clusters identified.
10. Click merge box beside each cluster and click merge selected and recluster.
11. Changed spelling of names, more clusters identified.
12. **Note:** clicked on the merge more than once, solutions for later exercises may be different.
13. Transforming data. Click items owned column, edit cells, transform.... New window opens. Can type general refine expression language (GREL).
14. Remove all left square brackets. In expression box type value.replace(”[”, “”), click ok. No longer left square brackets in item\_owned column.
15. Use same strategy to remove single quote marks, right square brackets in items\_owned column. Used the following GREL as directed in solution: value.replace("&apos;", ""), value.replace("]", ""), value.replace(" ", ""). All on separate lines. Mostly successful, but can still see occasional right bracket next to a value. Otherwise, all replaced with semi-colons.
16. Ran value.replace("]", "") for a second time on its own. Successful. All right brackets disappeared.
17. Use text facet to see what items were commonly owned. Click items\_owned, choose facet, custom text facet. Type value.split(”;”) in expression box. Click ok. Box appears on left panel with all items owned.
18. Which 2 items are most commonly owned? Sorted by count. Mobile phone and cow plough.
19. Which 2 are least commonly owned? Solar torch and solar panel.
20. Used same steps, cleaned the months\_lack\_food column. Got error at last custom text facet.**Error-** Left panel appeared saying: *Parsing error at offset 12: Missing number, string, identifier, regex, or parenthesized expression.* How to fix???? Not sure where I went wrong or how to recover steps.
21. Copied and pasted the commands from solution earlier to get exact typing right for the  months\_no\_water column. Successful.
22. Added GREL language. Text facet appeared on left.
23. Repeated cleaning and GREL steps for liv\_owned, res\_change, no\_food\_mtiigation columns. All successful.
24. Repeated GREL command using history tab. Clicked reuse next to command.
25. Using undo and redo. Click undo/redo on the left. All operations/commands are listed.
26. Clicked on a number of past commands to see the change in the dataset. Tried to see where I want wrong with the error in months\_lack\_food column but could not work it out.
27. Trim leading and trailing whitespace. Create new text facet for respondent\_wall\_type. Edited data so that spelling was correct and condensed identical choices.
28. Remove whitespace. Choose edit cells, common transforms, trim leading and trailingwhitespace. Only 4 choices in text facet in left panel.

Filtering and sorting with openrefine

1. Click on arrow next to respondent\_roof\_type, select text filter. Facet appears in left.
2. Type mabat, press enter. The column shows the rows with matching info. There are 58.
3. Change view to show 50 rows. Will see most of matching rows.
4. What roof types are selected? Mabatisloping and mabatipitched.
5. How would you restrict this to only one type? Be more specific with choices of letters/word in filter.
6. Excluding entries. Create text facet for respondent\_roof\_type. Drop down menu facet, text facet. Facet appears in left panel with 2 entries that agree with text filter.
7. Use include/exlude to select entries from one of these roof types. Hovered mouse over the entry and clicked include on mabatisloping. Other entry is excluded.
8. Sort. Select drop down arrow in gps\_altitude column. Select sort by numbers and smallest first. Column sorted in order. First few values are 0. This may be problematic, maybe missing data?
9. Clicked sort again, and selected reverse to reverse sort. Successful.
10. Sorting by multiple columns. Sort on gps\_longitude with number largest first. Successful.
11. Sort gps\_lattitude as number with largest first. Successful.
12. Select drop down arrow on village column. Select edit column, move column to end. Column moves to end so it can be compared easier with GPS coordinates.
13. Find village 49.  Can’t see what village it is based on coordinates.
14. Sort interview\_date column by data. Select sort, date.
15. Move village column to start. Drop down, edit column, move to beginning.
16. Row for 49 corresponds with interview times for Chirodzo village. Unsure about GPS coordinates.
17. Open text facet for village. Drop down arrow, facet, text facet. In left panel, select edit on 49. Change v49 to Chirodzo.