Learning Journal

Lauren Franks

FOAR705

Contents

1 Data Carpentry **1**

1.1 Formatting data tables in spreadsheets **1**

1.2 Dates as data **3**

1.3 Quality assurance **4**

1 Data Carpentry

Formatting data tables in spreadsheets

Exercise 1:

Objective: To clean up messy data sheets using Open Refine
Action: Opened OpenRefine software, uploaded SAFI Messy data set and created new project.
Error: Realised that I lack sufficient expertise in OpenRefine to proceed.
Result: Returned to excel.

Objective: Identify problems with data set
Action: Reviewed data carpentry literature for guidance and identified following errors: Colourful cells, misspellings, inconsistent recording styles, empty cells where there should be zeros, uses multiple sheets, uses multiple tables, uses symbolic formatting to show information, uses white space.
Result: successful identification of errors.

Objective: To clean up Mozambique data according to previously identified problems Actions:

1. Created new sheet in Excel and copy pasted data for Mozambique
2. Added new column for 'inc barn' and removed coloured cell

3. Moved 'livestock' and 'plots' graph to separate columns adjacent to dwelling graph

4. Created four separate columns for poultry, oxen, cows and goats

5. Created separate columns for 'water use' and 'water use summer only'

Errors identified in data:

- Key id 6 amount/type of livestock is not clear in original data set
- Meaning of 'look after cows' column is unclear in original data set as it seems to directly contradict information in original 'livestock owned and numbers' column which suggests that key id 2, 4, 5, and 9 all have cows
- Both livestock and plot data from key id 10 is omitted entirely from original data set

Result: Data is much clearer to read, however errors listed above are unable to be sufficiently addressed without further information from original researchers.

Objective: To clean up Tanzania data according to previously identified problems
Action: Copy pasted data for Tanziania into cleaned data tab on spreadsheet
Result: Success
Action: Created new column for the country to differentiate between key id locations in Tanzania and Mozambique
Result: Success
Action: Created new column for 'inc cowshed'
Result: Success

Errors identified in data:

- Unsure if 'burntbricks' and 'sunbricks' refer to the same value
- Key id 16 says it has 1 poultry, 1 oxen and 1 goat and yet a total of 4 livestock
- Key id 18 says it have a total of 1 livestock but does not list which kind
- No livestock data for key id 20
- No data for plots

Solutions:

- Added column 'livestock unspecified'
- I have opted to leave spaces black where data is unavailable as this is the best null value according to the Data Carpentry website.

Results:

- Saved sheet as a new excel spreadsheet called 'SAFI messy fixed Franks2019'
- Data is much clearer to read, however there remain some missing values as I was unable to deduce the answers from the available data. To fully complete data set would require discussion with original researchers.

Exercise 2:
Objective: Identify what types of metadata should be recorded from the regarding the data set.
Action: I discussed with Ellen about what types of metadata should be recorded. As per the Data Carpentry website, we concluded that some types of metadata that should be recorded and made available with the data are:

- the exact wording of questions used in the interviews (if interviews were structured) or general prompts used (if interviews were semi-structured)
- a description of the type of data allowed in each column (e.g. the allowed range for numerical data with a restricted range, a list of allowed options for categorical variables, whether data in a numerical column should be continuous or discrete)
- definitions of any categorical variables (e.g. definitions of "burntbricks" and "sunbricks")
- definitions of what was counted as a "room", a "plot", etc. (e.g. was there a minimum size)

Two examples of problem data produced in politics and international relations:

1. Adam Smith, 18th century philosopher and arguably the founder of modern economic theory, wrote that barter was a precursor to money. Despite there being no evidence for this claim it appears in many politics and economics textbooks. A similar false claim is 'trickle down economics', this flawed idea is used as data to back up flawed arguments.
2. Data bias in policing is a political issue.

1.2 Dates as data

Exercise 1:
Objective: To separate dates into components in a spread sheet
Action: Download and open dates excel file
Result: Success
Action: Created 3 new columns for day, month and year and separated dates into those columns
Result: Success

Action: Tried to used the automated functions on excel to input the days.
Error: It put a whole date in instead of the day.
Solution: I chose to enter the days manually into excel as there weren't too many, however, this is not an acceptable solution for the future and I will need to revisit this.

1.3 Quality assurance

Objective: Apply data validation rule to restrict data to a specific numeric range
Action: followed instructions on website and didn't have any issues.
Objective: Save excel file as csv file
Error: Cannot save csv with multiple worksheets
Solution: Saved both sheets as separate csv files.