Learning Journal- Thomas Kongonis

**Data Carpentry 01- formatting data tables in spreadsheets.**

Diagnosing issues

* Primary issues as advised in lesson plan is that the data-set includes 2 sheets and multiple tables per sheet.
* Solution: The obvious solution is to integrate all of the tables and sheets into one table on one sheet. Due to misunderstanding the expectations of the task, I proceeded to actually fix the data-set and implemented this exact means to fix the data-set.

**Dwelling tables (both Mozambique and Tanzania)**

* Utilising colour coding or asterisks to reference to external information outside of the tables. (The Mozambique data-set utilised the colour coding to refer to the inclusion of barn for key\_id 10. In terms of the Tanzania data-set, an asterisk was used in a similar fashion to refer to the inclusion of a cow barn for key\_id 8.)
* Solution: In both cases, the appropriate solution would be to remove the external information and remove the colour coding or asterisk reference to the information. This would be followed by creating two separate tables for the dwelling section of the integrated data-set. These two tables would be: inc\_barn and inc\_cowshed. This exact solution was implemented in my cleaning of the data-set. This was only possible after integrating both dwelling sections into the one table and applying these two new sections to all 20 of the key\_id’s which were then labeled 1-20 with Mozambique remaining as 1-10 and the Tanzania sample being referred to by 11-20. To remove any confusion surrounding this, directly next to the key\_id column, a country column was added.
* Titles of each table comprised of merged cells.
* Solution: a solution to this issue that I thought of and decided to implement in the attempted cleaned data-set was to remove the titles altogether and when including the data into the integrated table to utilise a inc\_dwelling, inc\_livestock and inc\_plot table respectively to both separate the data into their own individual sections and to remove confusion. This not only acted as a title for each section but also did so in a way that did not require useless data and the merging of cells.
* Spelling errors and poor formatting of data types.
* Solution: The obvious solution in this case was to fix spelling errors. (eg errth to earth) In terms of the format of the data types, the decision that was made during the data cleaning process that I undertook was to utilise camel case for the actual included data and to utilise the underscore key in lieu of a space to be in line with the key\_id column. During the process, I experimented with the multiple options suggested and admittedly settled on the options that I found most aesthetically pleasing hoping that it would help abate the slight screen headache I got!
* -99 input in the room section on the Mozambique dwelling table.
* Solution: This appeared to be an attempt at inputting a null value, initially I changed this within the cleaned data-set to be a blank cell but later decided to change it back. I did this as I decided that to actually change the data that was input without being able to consult those who took the data was an ethical issue. If I were to misunderstand their input and decided without the necessary consent and information it could cause a significant error within the reading of the table. As such, in the process of cleaning the table I decided to leave the data as is. There were other concerns similar to this but they will be addressed within a separate dot point.

**Livestock Table (Both Mozambique and Tanzania.)**

* Merged cell title.
* Solution: same solution as the dwelling table was utilised in this case.
* The recording of the livestock
* Solution: This section was quite possibly the easiest to diagnose but the most difficult to solve. Not only did each table have its own problems endemic to it, but to integrate the two due to this and to include all the original data was exceptionally challenging and I believe that my finished product is most certainly flawed in some ways that I was not able to identify. That being said, in the case of the Mozambique livestock table, the issues with the table are firstly the inclusion of numbers, words and multiple data inputs per cell. This left the data in an agonizingly frustrating state and made me realise how significant an understanding of this is to research. If I were to come across this in the research stage of a thesis and be totally unaware, its highly likely that a meltdown would be a certainty! To solve this multitude of issues all of the data needed to be made sense of which was in itself a flawed process. A new subsection of an integrate table was formed and there was a table included for each animal. Later, for the sake of including all of the data given, three more sections were added. These included: an animal\_unspecified table, a cow\_looked\_after table and finally a cow\_died table. The cow died table was added to include the data input for key\_id 3 in the Tanzania table that was input as yes/no and followed by an asterisk informing us of the cow’s death even though it was looked after. Following this, firstly the Mozambique livestock data was transferred to this new cleaned table. The Tanzania livestock data was also similarly transferred although this was easier due to the fact that this table only included one form of data per cell. Although this posed a question that I will ask in class: Is numerical and alphabet formats acceptable to use in tandem? I decided that it seemed messy and confusing so changed the yes to include a number that took into account the total table minus the numbers of the other livestock in their respective tables.

**Plots table (only provided for Mozambique)**

* Merged cell title
* Same solution used
* Data only provided for one country
* This does not appear to be a problem when simply diagnosing issues and possible solutions yet in practice and action, new problems arise that theory cannot account for. As such an inc\_plots table was incredibly useful when integrating both data-sets into one table. Furthermore blank cells were utilised to signify null value.
* Poor data recording
* Solution: Some inputs on further assessment were actually null values that could be correctly diagnosed as such. For example key\_id 5 had an empty cell for plot number and an n was recorded in plots watered. This was enough information for me to comfortable list that this be listed as a No under inc\_plots. Key\_id 10 was clearly null and not included however key\_id 9 was also added as null because of the ridiculous way it was recorded. -999 is clearly a null value and a 1 was added for a yes or no section which is why it was safe to list it as not included. Key\_id 8 was more difficult to solve. The solution to this was to include a plots\_watered section as the data seemed to suggest that a 1 in this case for water use was only for one plot out of two. In terms of key\_id 3, another row was added to account for water use in summer and the water use table was replaced with a continual water use row in which the yes values were included as yes in the continual section and yes in the summer section.
* Tanzania table listed as data collected in 2017
* This was information included within the table and as such another row was added at the end called year\_collected. In the Mozambique section this was left as null and 2017 was added to the Tanzania section.

**This information is all a distillation from the diagnosis and data cleaning process that was implemented on the table as was my mistake. As such, the journal recordings made when the data cleaning was undertaken will be included under this paragraph for transparency and the cleaned table will also be submitted for the sake of feedback and to showcase if there was anything I missed and forgot to record within the preceding part of the journal.**

* Diagnosed problems
* Printed spreadsheets to troubleshoot due to headache from looking at screens, this gave me a bit of a break whilst continuing work
* Begun cleaning the Tanzania sheet:
* Initially made copy to begin working.
* Firstly: dwelling table re named all of the values to a constant- implemented camel font whilst retaining underline for the titles.
* Changed key id 3, room to null value as the -99 seems to suggest that a measurement was not taken
* Capitalised all values that did not have a second word for continuity.
* Figured out what the colour coded square meant. This was to suggest that this house had a special room or barn and this accounted for the extra room.
* Decided that it would be more effective to create a new table for barn.
* Implemented this with the values being either ‘Yes’ or ‘No’ and removed colour code and separate value external to the table.
* Encountered problem, could not get black outlines on this new part of table, left this to address if the problem is found again.
* Created new sheet to include the fresh and all-inclusive table.
* Listed Key ids 1-20 and assigned another table to signify country.
* Decided not to merge any cells to make it more exportable and readable.
* Instead decided to add another row titles inc\_Dwelling and signaled Yes for all inputs.
* Copied corrected roof\_type inputs for key\_id 1-10 and added them into new table.
* Proceeded to do the same for wall type floor type and rooms and added inc\_barn table.
* Moved to Tanzania dwelling table for cleaning.
* Proceeded to format in same way as Mozambique.
* Diagnosed cowshed problem that was similar to barn problem, except no colour and asterisk was used instead.
* Implemented same solution- implemented inc\_cowshed in new table.
* Utilised search function and implemented freeze top row to increase ease.
* Solved border problem by searching through format cells menu.
* Set up animal section of the table, rather than merging cells, decided to create an inc\_animals section.
* Decoded awful table in Mozambique section, had difficulty and had to redo after realising that there were some unspecified animals, this was the same for the Tanzania section.
* Added null sections for key id 10 and 20 as they were not measured, used null values because all the ones and zeros confused me, will remove later.
* Moved to plots table.
* Created inc\_plots section for clarity.
* The confusion in this table lied a lot in all the unsuspecting null values, the decision was made to treat these as null values, this included id 6,9 and 10
* Created a constant water use and summer water use value sections. Id 3 only used water in summer.
* As for 7 and 8, their water use was constant so the decision was made to make their summer water value yes.
* Id 8, in the original table had the water value as one, the decision was made to include a section that states amount of plots watered and this was set to 1, the other ids that had continual and summer watering were set to include all the plots.
* Table was formatted and given lines.
* Upon review a mistake was seen, looked after cows’ section was ignored, this was inserted at the end of the animal section. The dead cow value was signified as yes.
* Animal death section was added to account for this data
* Date collected section added to include all data
* Only available for Tanzania.

**Lesson 02- Metadata**

When reading through the SAFI\_clean table, besides being still slightly confused by spreadsheet data inputs, there appeared to be several things that stood out that i felt would require metadata in a separate document to help clarify. These included:

* The countries each of these towns are in, its possible that this is all the same country and i wasn&apos;t sure if there was continuity from the last spreadsheet looked at. That being said i thought it would be useful to list the country/ies that these towns are in and possibly their proximity to eachother.
* A second thing that stood out to me in both tables was the types of bricks, it would be useful to know what they are comprised of and how well they fulfill their ‘brickness’.
* Thirdly, what exactly connotes a room would be useful. By this i mean is there a minimum or maximum size.
* Fourthly, a big concern i had was: what justifies items\_owned? Is there a monetary amount or a necessity for a large purpose? For example, a shirt could fall under this category so it would be good to know what the lower range is for inclusion.
* Fifth, what constitutes a meal would be necessary metadata. We all have very different notions of what a meal is. Is a full belly necessary or is just staving of hunger enough?
* Sixth, does months lacking food constitute and amount? For example, would one meal a week class as lacking food or having food for the month?

**Scoping exercise:**

* I made this exercise more difficult for myself by starting with paper, moving to word then finishing with LateX. I was scared of the unknown but in reality LateX was pretty cool and i’ll be using it going forward.
* I begun by utilising the information given to us to create a funnel. I then began to write it all down and transcribe to word as i realised that i did a few things wrong and edited. (misunderstood some of the distinctions between the categories)
* During this process i had an idea for the final assessments that i noted and added into the exercise.
* After this i moved to Overleaf, after creating my account and selecting a new document i saw the rich text section which i begun to utilise first.
* I proceeded to write down titles and then i made bullet points and copied and wrote down the exercise in full.
* After i finished inputting the information i then decided to go to the learn LateX page as it all seemed too simple and i was surprised that i had made a document without breaking it.
* I then proceeded to read and realise what all the code was for. I decided to play around with some of the code to show that i made an attempt to learn.
* Initially i saw that the title was made from code and it utilised my name as it appeared in my student email. I altered this and moved my name and student number into the author section and kept the \title section as: Scoping Exercise.
* I then saw that there was code for sections and sub sections which i decided to play with. I took the Jobs, pains, pain relief, gains and gain enabler sections and coded headings in rather than just typing. Then as i continued to read i saw the code for \tableofcontents. I decided to get this working. So i added the code into the start of the document below the title of the document.
* I then proceeded to change the code for the headings to \secton{}.
* After i updated the document, i realised i had made a mistake, the table of contents was there but the title and author was no longer.
* To fix this i attempted to take the begin document function and add it above the title.
* After this didn&apos;t work i deleted the begin document function and retyped it. this then meant that the end document function was just below so i broke the document and it all disappeared.
* I realsied this problem and fixed it but the title still wasn&apos;t there. I referred back to the site and realsied i was missing a \maketitle function for some reason. I input that and then it was all fixed.

* I then saved the document as a pdf and a tex file, uploaded both of these to a folder in cloudstor and just uploaded the pdf to Ilearn.

**Dates in Spreadsheets:**

* Begun by copying dates into a new sheet and proceeded to input values
* Inserted values through utilising the codes and got bad results
* Trying to input the days was fine, however inputting the month as =MONTH(11) returned a value of 1. Tested this in the year section and the =YEAR(2016) function returned the result 1905.
* Upon reviewing the lesson again I realised that the mistake was that it was utilising the number date system and the number 2016 implied 2016 days from December 31st 1899. The solution that I found to this problem was by inputting the date, year and month code and linking the date column to the code for example =Date(A2). After applying this to the entire table I got the correct results.

**Second part:**

* Followed instructions, the first column was changed to say Nov-17 then changed the year to 2017.
* Realised that the version of excel I was using utilised American dating. As it populated the new data point because there was no year it instantly assumed that it was current year so showed 2019.

**Quality Assurance lesson:**

* Followed instructions and input the minimum and maximum numbers for the No\_membrs column to 1 and 30 respectively.
* Created an input message for the Room column that included a min of 1 and a max of 10 as this seems to be fair, any house with more than 10 rooms would probably be a mansion!
* The input message stated that the column must not exceed a max of 10.

List selection:

* Followed first example, decided to apply this to the Town column. There was only 3 towns so this made the list option very fitting for this column. I included the correct spelling by following what has already been written and left a meaningful comment that stated my intention and that any new towns added to the data set must be first approved and then added to the dataset through the data validation tools.

Exporting lesson:

* Utilised the .csv file type to save the copied SAFI clean spreadsheet.

Scoping II

* Experimented further with coding in Overleaf.
* Created sections and subsections with the relevant code.
* A difficulty faced however was no matter what I tried to create a space between text without going to a new section, I could not. I went through information on the overleaf site and found a code that could be implemented, this was the \vspace code. I set this to 5mm and achieved success with what I wished to happen. This allowed for me to split up the writing as I wished.
* I then proceeded to apply this to the sections and subsections so that I did not need to create constant new lines of code and over complicate the document.
* Finished inputting necessary information and backed up the document ready for upload.