

Learning Journal

Emily Hunt

August 2019

1 LaTeX

1.1 Objective: Create a section

Time Stamp 10:52am 12 August

Actions: In the source tab enter `\section{section title goes here}`

Results: The numbered section was created successfully.

Errors: None.

Solutions/Notes: For unnumbered sections, include an asterisk before the `{section title}`. For a subsection, replace `\section` with `\subsection`.

1.2 Objective: Bold some text

Time Stamp: 11:00am 12 August

Actions: In the source tab enter `\textbf{text you wish to bold goes here}`.

Results: The text was put into bold successfully.

Errors: None.

Solutions/Notes: To underline text replace `\textbf` with `\underline`, and use `\textit` for italics.

1.3 Objective: Create a line break in a block of text.

Time Stamp: 11:10am 12 August

Actions: Insert `\\` at the end of the line I want the space at.

Results: The text was placed on a new line.

Errors: At first using the two back slashes prompted the error - Underfull `/hbox (badness 10000)` in paragraph at lines 15-19.

Solutions/Notes: I removed the `\\` and instead inserted an extra space between the lines in the source tab using the enter key. This seemed to work successfully and removed the error. I then went back later and tried again with the two back slashes and this no longer produced an error so I reinserted them where I wanted the line breaks.

1.4 Objective: Align text to the left.

Time Stamp: 11:20am 12 August

Actions: I inserted `\usepackage[document]{ragged2e}` at the top of the document.

Results: This aligned the entire text the left.

Errors: I received a note in the log the command had changed.

Solutions/Notes: I put `\usepackage{ragged2e}` at the top of the document and placed `\begin{FlushLeft}` at the start of the text I wanted to align and `\end{FlushLeft}` at the end of the document, and this seemed to work.

1.5 Objective: Include Code in the Text.

Time Stamp: 11:35 12 August

Actions: I wanted to allow some parts of the code to be seen in the text so I could properly note my actions, and just including them would execute the command. I started by using the listings package, so included `\usepackage{listings}` at the top of the document, and put `\begin{lstlisting}` before the code I wanted to be visible and `\end{lstlisting}` after.

Results: This worked successfully.

Errors: However, while not an error, it placed the code on a new line and in a larger font which made the document difficult to read.

Solutions/Notes: I then tried using `\begin{verbatim}` and `\end{verbatim}` around the code I wanted, but this had a similar look to using the listings package. Then, I tried the `\verb` code, and this had the effect I desired.

1.6 Objective: Create an Bullet List

Time Stamp: 3:00pm 12 August

Actions: I put `\begin{itemize}` at the beginning of the list I wanted to create, `\item` for every bullet point, and `\end{itemize}` at the end of the list.

Results: This worked successfully.

Errors: No errors.

Solutions/Notes: For an ordered list change `{itemize}` to `{enumerate}`

1.7 Objective: Sync my Learning Journal on Overleaf to GitHub

Time Stamp: 8:00pm 12 August

Actions: I went to the Menu Tab, clicked on GitHub under sync, clicked 'Link to GitHub', signed into my GitHub account and authorised OverLeaf.

Results: My GitHub and OverLeaf accounts were linked.

Errors: Going back the menu and clicking GitHub, it says the document is not linked to a GitHub repository, and that I need to create one. I would like to link it to my existing repository created in class.

Solutions/Notes: I need to work out how to put it into my existing repository.

1.8 Objective: Insert a clickable link

Time Stamp: 8:10pm 12 August

Actions: Add `\usepackage{hyperref}` to the preamble. To insert the link, include in the source, `\url{link goes here}`.

Results: Success link created

Errors: None

Solutions/Notes:

2 Data Carpentry

2.1 Introduction

Have I Used Spreadsheets in My Research?

So far throughout my time at university, I have not used a spreadsheet in my research.

Have I Done Something that Made Me Frustrated or Sad?

I have definitely done things that are frustrating, like not being able to find references I have used and accidentally leaving some out in my final submission. Also, formatting bibliographies and my research can be a frustrating process.

2.2 Formatting Tables in Spreadsheets

Activity 1: What is Wrong with the Messy Data and How Can It Be Fixed

Mozambique

- The absence of an observation is indicated in a number of different ways e.g. -99, -999 or cells are left blank. This could be addressed by choosing one and using it consistently
- In the Dwelling table, there are a number of differences in spellings and identifications between the observations e.g. mabati-sloping v mabatis-sloping. One should be chosen and consistently used.
- In the Dwelling table, the yellow fill of the cell to indicate the inclusion of the barn will not be able to be processed by the computer during analysis. Potentially the ownership of a barn could be included as a separate variable.

- In the Livestock table, the livestock-owned-and-numbers column contains two variables. To address this, it could be split into observations for each individual animal, as exemplified in the example.
- In the water use column of the Plots table, there are data switches between ‘no’ and ‘yes’ and Y and N. One should be chosen and consistently used.
- Also in the water use column of the Plots table, there is an extra comment for key-id 2. This should either be entirely removed, or separate columns set up for the different seasons.

Tanzania

- Similarly to the in the Mozambique Dwelling table, the use of an asterisk to indicate the inclusion of a cowshed is not an appropriate way to label the data. Cowsheds could be treated as a separate variable.
- This is also the case in the Livestock table, where key-id 3 has ‘yes/no*’ in the Look after cows column. Either yes or no should be chosen, or an observation committed completely.
- In the Livestock table, the data for key-id 5 is recorded with yes and no, as opposed to the numerical data primarily used for the other responses. This should be changed into quantitative data.
- There are a lot of blank cells in the Livestock table. A single indicator of a lack of response should be decided upon and used consistently.

Generally

- The data set for Tanzania is missing data on Plots. Headers that have more than one word switch between having a space between them and an underscore.
- A single format for the headings should be consistently used.
- The two tables recording Livestock are set up differently. Also, one table records responses for poultry as either yes or no, while the other records them numerically.
- These issues could be addressed by fixing the Mozambique Livestock table as previously discussed.
- Across the two countries, the key-ids consist of the numbers 1-10. When the two country’s data is merged, they would be understood to be the same people. To address this, Tanzania’s could be changed to 11-20.
- Livestock data for key-id 10 is missing from both tables.
- All of the data should be placed into a single table

Activity 2: What types of metadata should be recorded with this project?

- What does NULL indicate (was the question not asked, did the respondent not want to answer etc.)
- Definitions of the different variables. For example, what constitutes a mudduab wall type, or what is the basic definition of a room.
- What were the exact questions asked in the survey to gather the responses.

2.3 Formatting Problems

No activities

Examples of Problem Data in my Discipline (Film Studies)

Problem 1: Data Spread Across Numerous Data Sets

Data Sets: <https://www.imdb.com/interfaces/> and <https://archive.ics.uci.edu/ml/datasets/Movie>

The IMDb data is split into a number of datasets, each with different headers. For instance, the title.akas.tsv.gz data set has columns for the title, region for the version of the title and language of the title, whereas the title.basics.tsv.gz data set includes columns for the type/format of the title, primary title and original title.

This is similarly the case with the University of California Irvine Movie Data Set, which has various data sets for 'People', 'Casts', 'Actors', 'Remakes' and 'Studios'. Additionally, in the various data sets, there are a number of columns that contain two variables. For example, in the Remakes data set, in the title and priortitle columns, there is a 'T:' in front of every entry to indicate it is a film title.

Though discussed vary vaguely, data from the IMDb database was used for

- Sehwan Oh, Hyunmi Baek and JoongHo Ahn (2017) Predictive value of video-sharing behavior: sharing of movie trailers and box-office revenue, *Internet Research*, 27:3, 691-708, DOI 10.1108/IntR-01-2016-0005.

Problem 2: Null Values

Data Sets: <http://www.cinematics.lv/database.php>

In the Cinematics Database, there are a number of cells left blank in a number of the entries and it is unclear why these cells have been left blank e.g. is the data unavailable or is there its not applicable?. Additionally, there are a number of repeated entries (such as for the film *Insidious The Last Key*).

With the database containing the average lengths of shots in films, a potential problem in the raw data is that it is susceptible to human error, as it is the length of a shot is manually measured by a viewer using a software that works like a stopwatch.

This misidentification of null values is also present in the previously mentioned University of California Irvine Movie Data Set. Incomplete data is identified in a variety of ways, including cells being left blank, and question marks.

The data from the Cinemetrics database appears in

- James E. Cutting, Kaitlin L. Brunick and Jordan DeLong (2012) On Shot Lengths and Film Acts: A Revised View, *Projections*, 6:1, 142-145, DOI: 10.3167/proj.2012.060106
- Lev Manovich (2013) Visualizing Vertov, *Russian Journal of Communication*, 5:1, 44-55, DOI: 10.1080/19409419.2013.775546

2.4 Dates as Data

???

2.5 Quality Assurance

???

2.6 Exporting Data

???