2024-02-07-lmu-online

Welcome to The Carpentries Etherpad!

Users are expected to follow our code of conduct: https://docs.carpentries.org/topic_folders/policies/code-of-conduct.html

Schedule

Please see all notes on the workshop page https://marionwalton.github.io/2024-02-07-lmu-online/

Break times (approximate):

00:50-01:00 (9:50am-10.am Pacific)
02:00-3:00 (11.00am-12noon Pacific)
03:50 - 4:00 (12.50pm-1pm Pacific)
You can keep track of the time in your current timezone at https://timeanddate.com/worldclock.

Sign in: Name (Pronouns), Institution, Email & Twitter/BlueSky/Mastodon (optional)

- Marion Walton(she/her) University of Cape Town marion.walton@uct.ac.za https://hci.social/@Marionwalton
- Lyrric Jackson (she/her) Spelman College, ljacks@spelman.edu
- Teresa Hebron (she/her), University of Utah, teresa.hebron@utah.edu, @EndivesArchives (Twitter/BlueSky)
- Ruth Monnier (she/hers), Pittsburg State University, rmonnier@pittstate.edu, @AMsRuth on Twitter
- Jesse Randolph (she/her), Santa Clara University, jrandolph@scu.edu
- Michele Gibney (she/her), University of the Pacific, mgibney@pacific.edu
- Tessa Withorn (she/her), University of Louisville, tessa.withorn@louisville.edu, @tessawithorn
- Alex Dryden (he/him), University of Illinois at Urbana-Champaign, adryden3@illinois.edu
- Simon Robins (he/him), University of Dayton, srobins1@udayton.edu, @SimonRobins10
- Elaine Yeung (she/her), Chapman University, eyeung@chapman.edu
- Kate Lambaria (she/her), University of Illinois Urbana-Champaign, lambari1@illinois.edu
- Lorna Dawes (she/her) University of Nebraska-Lincoln. ldawes2@unl.edu
- Gina Cherundolo (she/her), Wilkes University, gina.cherundolo@wilkes.edu
- Taylor Parks (they/them), Loyola Marymount University, taylor.parks@lmu.edu
- Rachel Smiley (she/her), California Lutheran University, rsmiley@callutheran.edu
- Peter Rolla (he/him), Loyoloa Marymount University, peter.rolla@lmu.edu
- Monica Carroll, UIUC
- Savannah Kelly (she/her), University of Mississippi, slkelly@go.olemiss.edu
- Stacy Katz (she/her), Lehman College, stacy.katz@lehman.cuny.edu
- Maria Waselus (she/her), University of Michigan, mwaselus@umich.edu
- Danielle Worthy (she/her), Loyola Marymount University, danielle.worthy@lmu.edu
- Nick D'Andrea (he/him), Pepperdine University, nicholas.dandrea@pepperdine.edu
- Chris Marino (she/her), Hoover Institution Library & Archives Stanford University, cmmarino@stanford.edu

- Steph Gritz, Loyola Marymount Univeristy, sgritz@lmu.edu
- Andrea Hebert (she/her), Louisiana State University, ahebert@lsu.edu
- Wen-ying Lu, Santa Clara University, wlu@scu.edu
- Sarah Follmer (she/her), University of Illinois at Urbana Champaign (sfollmer@illinois.edu)
- Jamie Hazlitt (she/her), Loyola Marymount University, jamie.hazlitt@lmu.edu, @jamiehaz (bluesky)
- Dominique Dozier (she/her), Santa Clara University , ddozier@scu.edu, instagram.com/dominique_sculibrarian ddozier@scu.edu
- Cameron Schwing, University of Illinois Urbana-Champaign, cschwin2@illinois.edu
- Tiffany Garrett (she/her), Roseman University of Health Sciences, tgarrett@roseman.edu
- Angie Rathmel (she/her): University of Kansas, aroads@ku.edu | atruthbrarian@gmail.com
- Kristin Laughtin-Dunker (she/her), Chapman Univeristy, laughtin@chapman.edu
- lara University)he/her),
- Grace Kim (she/her), Nevada State University, grace.kim@nevadastate.edu
- Cat Hannula, (she/her), Amherst College, channula@amherst.edu
- Michelle Ehrenpreis (She, Her) Lehman College, CUNY, michelle.ehrenpreis@lehman.cuny.edu
- Glenn Johnson-Grau (he/him), Loyola Marymount University, gjohnson@lmu.edu
- Sanjeet Mann (he/him), CSU San Bernardino, sanjeet-singh.mann@csusb.edu
- Heather VanDyne (she/her), Fort Hays State Unviersity, hpvandyne@fhsu.edu
- Amber Janssen (she/her), California State University, Maritime Academy, ajanssen@csum.edu
- Melissa Gomis, University of Nebraska - Lincoln libraries, melissa.gomis@unl.edu (she/her)
- Ashley Wilson (she/her) San Diego State University, aswilson@sdsu.edu
- Jessica Serrao (she/her), Clemson University, jserrao@clemson.edu
- Gabriel Foster (he/they), University of Illinois, gtf2@illinois.edu
- June Kim, National University, jkim3@nu.edu
- Lorelei Rutledge (they/them), University of Utah
- Andrew Greenman (he/him), Chapman University, agreenman@chapman.edu
- Courtney Block (she/her) Indiana Univ Southeast coblock@iu.edu
- Catherine Baird, she/her, Montclair State University, bairdc@montclair.edu
- Jonathan Mccormick (he/him) Gateway Seminary - JonathanMcCormick@gs.edu
- Emily Croft, University of Redlands, emily_croft@redlands.edu
- Sarah Parramore, Occidental College, parramore@oxy.edu
- Daniel Keough (he/him), Hoover Institution Library & Archives - Stanford University, dkeough@stanford.edu
- Jennifer Embree (she/her), Binghamton University, jembree@binghamton.edu
- Samaa Haniya (She/Her), Pepperdine University, samaa9000@gmail.com
- Colleen Mullally (she/her),Pepperdine University
- Bradley Dusing (he/she/they), Trinity University, bdusing@trinity.edu
- Sarah Christensen (she/her), University of Illinois Urbana-Champaign.
- Sarah Lopez (she/her), United States Naval Academy, lopezs@usna.edu
- Kara Kugelmeyer (they/she), Colby College, kara.kugelmeyer@colby.edu

Stephanie Ward (she/her) - University of Northern Colorado - stephanie.ward@unco.edu
Jessea Young (she/her), Loyola Marymount University, jyoung46@lmu.edu
Denise Quintel, she/her/hers, Middle Tennessee State University, denise.quintel@mtsu.edu
Shu Qian, Worcester State University, sqian@worcester.edu
Ana Munandar, Chapman University, munandar@chapman.edu
Astra Yatroussis (they/she), Loyola Marymount University, astra.yatroussis@lmu.edu
Leslie Delserone (she/her), University of Nebraska-Lincoln, ldelserone2@unl.edu
Rebecca Orozco (she/her), United States Air Force Academy, rebecca.orozco@afacademy.af.edu

Brooke Troutman (she/her), United States Air Force Academy, brooke.troutman@afacademy.af.edu
Erin Durham Wright, (she/her), University of Maryland Baltimore County- UMBC, edurham@umbc.edu
Cristina Vasquez, (she/her), Loyola Marymount University, cristina.vasquez@lmu.edu
Kelly Durkin Ruth (she/her), United States Naval Academy, durkinru@usna.edu
Amy Pham (she/her), University of San Diego, abpham@sandiego.edu
Maja Subelj, UA
J. Denice Lewis (she/her/hers), Wake Forest University, lewisjd@wfu.edu
If you have a moment before we begin and have not yet done so, please fill out the pre-training survey at https://carpentries.typeform.com/to/wi32rS?slug=2024-02-07-lmu-online

Welcome activity

Why are you taking this course? What goals do you have for this training?

- Teresa H: learning better techniques for manipulating/managing research data for a current project; brush up on existing skills
- Ruth M. - learn more skills to assist in research/upskill (very weak on data science skills) (100% to Sanjeet)
- Chris M: tools to assist in data clean up and analysis for researcher and job related statistics reporting
- Simon R: learning new skills for data cleaning, processing, and visualizing.
- Love to learn new stuff about data
- I am a Data Services Coordinator and very new to the role so this training will be so helpful for me as I dive into the world of data. I am also exploring data analysis in my current research project. My goal is to come away with more confidence in data management and analysis methods.
- Stacy K: I started learning R and data cleaning in my graduate studies and I'm interested in learning more and expanding/reinforcing this knowledge
- Alex D: Bush up on data management skills/systems, observe/learn techniques for teaching data skills
- Bradley D: I am interested in learning how to interpret social science survey data
- Elaine Y: refreshing my R skills and learning new data management methods
- Sanjeet: I've been wanting to learn OpenRefine and R for years but never set aside the time - learning together in this workshop is a great way to get started! Also excited to participate in the Carpentries.
- Jonathan McCormick: I oversee doctoral students and instruct instructional design. I am wanting to improve my ability to assess student outcomes.
- learn new skills for my own research and to share with students on my campus
- Learn if there are better ways to do things
- To refresh and learn some new tools/skills for data analysis.
- To better manage data
- Better management of data
- To learn new skills, be introduced to things I haven't thought of learning
- Andrea H.: learning how to clean data more efficiently
- Kate L: To learn new skills and refresh some old ones to improve data management
- MElissa Gomis: To relearn some skills and tools I used to know and to learn how to work with my data more efficiently.
- Nick D'Andrea: To expand my understanding of data and analysis

- Sarah Christensen: I'm a helper trying to expand my knowledge base
- Sarah Follmer (UIUC): Looking forward to getting some experience with OpenRefine!
- Lorna D. Introduction to R and review OpenRefine to manage research data.
- Rachel S: I'm hoping to learn new data skills to assist me in organizing and analyzing library data across the department. I'd love to set a goal of creating institutional reports for my campus.
- To learn new skills and refresh existing skills surrounding data management, analyses, organization etc.
- I need more tools for data analysis
- Peter Rolla: I have taken previous Carpentries courses and really liked them. I am hoping to learn more about the tools I already know a little of and to learn new tools
- Jesse R.-- I learned R and data curation in my MLIS but boy has it leaked. So this is refreshing those skills, and learning OpenRefine, which I've been hearing more and more about.
- Michelle Ehrenpreis- I want to improve my data organization skills and data analysis skills as well
- Courtney B: to learn new tools and skills
- To brush up on my skills and learn new skills to improve my current data management and analysis process
- Colleen M: I would like more exerience using organizational tools and brush up on OpenRefine.
- Jamie H - this is new to me, and I'm working on a grant that may have more complex data organization needs than I have experienced before - day 1 seemed like an accessible place to start!
- Stephanie W. - This is all new to me and I just want to become more familiar with data management/analysis.
- I know how to use SPSS but learning R always seemed daunting to me.
- Denise Gehring: Learn new skills
- Heather V: I want to learn new skills and tools to manage data more effectively
- Sarah L- brushing up on skills
- Maja Subelj-Improving the skills
- Daniel K: Feeling comfortable with large data sets and learning new ways to manage/visualize/manipulate data
- Need help managing and analysing research data.
- John Paul Odiegwu
- Help students more effectively
- To learn new skills and knowlege
- Ana M: learn new skills.
- Astra Y: Expand my data analysis skills early in my career
- Cat H: Introduction to R!
- Catherine B. - Have tried once before but need more to get R figured out.f

Dominique: After reading the article, I just want to develop "good enough practices" to get started!
https://excalidraw.com/#room=cb82596d389eda83ebf4,2_2Oy8R1L6-_eZYA8RDCeA
Angie: Refreshing and learning. Have used OpenRefine, not R. Use it in regualr work and want to explore
application to research.
Leslie D.: Using R and OpenRefine but not in the context of social sciences data
Rebecca O: Learning more about data management tools

Day 1 Feb 7, 2024

Helpers for today: Jessica Serrao, Savannah Kelly, Sarah Christensen, Courtney Block, Simon Robins | 9am-2pm Pacific | 10am-3pm Mountain | 11am-4pm Central | 12 noon-5pm Eastern |
Session 1 Data Organization in Spreadsheets (Marion Walton)

- 0:00 - 0:15 Introductions and icebreaker
- 0:15 0:30 Formatting Data Tables in Spreadsheets
- 0:30 - 00:50 Formatting Problems
- **00:50-01:00** Short break
- 01:00 - 01:20 Dates as Data
- 01.20 - 01:40 Quality Assurance
- 01:40 - 02:00 Exporting Data

Long break: 02:00-3:00

Session 2 Open Refine for Social Science (Lyrric)

- 03:00 - 03:15 Introduction
- 03:15 - 03:50 Working with OpenRefine
- **03:50 - 4:00 Short break**
- 04:00 - 04:30 Filtering and Sorting with OpenRefine
- 04:30 - 05:00. Examining Numbers in OpenRefine

The SAFI data

Farmer-led irrigation development in Africa on Youtube (Mozambique)
https://www.youtube.com/watch?v=agNxd7l0tyQ
More info about the data https://www.datacarpentry.org/socialsci-workshop/data
CITATION: Woodhouse, Philip; Veldwisch, Gert Jan; Brockington, Daniel; Komakech, Hans C.; Manjichi, Angela; Venot, Jean-Philippe (2018): SAFI Survey Results. doi:10.6084/m9.figshare.6262019.v1

If you haven't done so yet, download the following three files of example data (it's not the full SAFI dataset):
SAFI_clean.csv https://ndownloader.figshare.com/files/11492171
SAFI_messy.xlsx https://ndownloader.figshare.com/files/11502824
SAFI_dates.xlsx https://ndownloader.figshare.com/files/11502827
Place these 3 files in a folder you can easily find and access on your computer (for instance in a datacarpentry-spreadsheets folder on your Desktop or within your Home folder).

Introduction to Data Organization

Spreadsheets

Have you used spreadsheets in your research? (Type Y or N here to answer) Y
YyYYYYyyyYYYYeYsyYyy

YYYYYYYYYY
YYYYYYYYY
YY
Y
Y
Y
Y
y
NN
Very little+1
Y

Have you accidentally done something with a spreadsheet that made you frustrated or sad? What was it?

- Marion: Made an error about sampling which meant that my student had to repeat an entire analysis
- Inserted a column on a filtered sheet and the new Column A wasn't filtered for a few weeks
- Sorted by one column rather than whole spreadsheet and data got all messed up - fixable! but still annoying
- Savannah: Used highlighting in an Excel file, saved it as a csv, and then lost all the formatting. Whoops! - this has happened to me before too!
- Jamie - ISBNs in spreadsheets make me frustrated and sad! +1 (agree - and also call numbers, sort?) +1 + And Alma MMS IDs! They are too long! and barcodesISBNs in general make me frustrated and sad!!
- Teresa Accidentally broke a VLOOKUP in an inherited spreadsheet that helped standardized metadata across sheets within the workbook
- Merged cells and made analyses much more difficult
- Leslie: Overwrote data!
- Definitely done/felt what others are sharing but I also feel like I have to revisit data spreadsheets often to add more data, to reorganize data, etc.
- Elaine: struggling with data comparisons using VLOOKUP
- Sanjeet: Tried to apply a counting formula in a column but the total didn't seem correct, had to go back through the column line by line to find out where the formula went wrong. Also, I would love to know how to use VLOOKUP and HLOOKUP, I've looked at the Excel help several times but never quite figured it out.
- Basic arithmetic error in my formula. If I was good at math I would have been an engineer!
- I feel like I have to google every little thing... +1 +1
- tried to convert sheet with hidden columns to CSV
- Connecting data (attempting) across sheets & forgetting to freeze the header after sorting of columns and data
- Accidental duplicates.
- Not cleaning up the data on the master tab and then deleting columns later where I should have copied the master file and deleted the columns and then made analysis tabs (made standardizing the columns more difficult later across the tabs)
- Excel insisted that one date was another date after it automatically converted a date range. Just had to carefully undo past where it made the decision, still no idea what happened. Also, made it fail to insert commas when exporting to .csv

- Optimist: The glass is ½ full. Pessimist: The glass is ½ empty. Excel: The glass is January 2nd.:):):):)LOLXDPassword protecting a sheet and losing the password!
- Rebecca O: having to resort or rearrange the spreadsheet to make the visual come out correctly

I can't think of specifics, but *frequently* I start editing and using formulas and realize I have a problem that I can't fix, and so the only option I can come up with is to close the spreadsheet without saving and starting again completely from scratch (Peter)+1. I don't know what I don't know! Except a hunch that there MUST be a better way.
sorting by column but not the spreadsheet and not realizing it until much later :(+1
error in formula
overwriting raw data
+1 savannha
+1 ISSN and ISBN issues
Incorrect formula that was difficult to correct
+1 to sorted by one column rather than whole spreadsheet and then things got weird!

Formatting Data Tables in Spreadsheets ("Tidy" data)

The cardinal rules of using spreadsheet programs for data:

Put all your variables in columns - the thing you're measuring, like 'household_members' or 'years_farming'.
Put each observation in its own row.
Don't combine multiple pieces of information in one cell. Sometimes it just seems like one thing, but think if that's the only way you'll want to be able to use or sort that data.
Leave the raw data raw - don't change it!
Export the cleaned data to a text-based format like CSV (comma-separated values) format. This ensures that anyone can use the data, and is required by most data repositories.

Exercise - How would you tidy the SAFI data?

We're going to take a messy version of the SAFI data and describe how we would clean it up.

Download the messy data.
Open up the data in a spreadsheet program.
Notice that there are two tabs. Two researchers conducted the interviews, one in Mozambique and the other in Tanzania. They both structured their data tables in a different way. Now, you're the person in charge of this project and you want to be able to start analyzing the data.
With the person next to you, identify what is wrong with this spreadsheet. Discuss the steps you would need to take to clean up the two tabs, and to put them all together in one spreadsheet.

Make note of your answers here

- Ruth/Dominique -> Key-Id as one column and therefore only one table; missing variable - when was data collected (Cow died in April?); Cowshed/Barn - same thing? and might need a new column; total column for livestock title; how do you have -99 rooms and -999 Plot? (data entry

error); What is a Plot? - plot usage variable (summer response); Water usage in/on Plot - yes, no, and 1s as entry - therefore we don't know the story (water on the plot like a pond OR official plumbing); "errth" typo;

- Angie/Tessa - Error in data maybe with -99 and -999 Answer: -999 is indicating missing data, better to use null or na to indicate missing data. Format livestock from Mozambique as it is structured in Tanzania. Remove annotations and put in a ReadME file. Use underscores for compound words. Fix spelling errors. Clarify if data is missing.
- Kristin/Colleen -- variable names inconsistent/misspelled in Mozambique dwelling set, negative number of rooms, duplication of key_ids if both data sets combined, multiple variables in one column for Mozambique Livestock data, Mozambique has "Plots" data set that Tanzania did not collect (?), some Tanzanian variables have asterisks for further explanation, Tanzanian Livestock has yes/no variables instead of numerical values for some of the observations
- Michelle/Shu-separate out the look after cows from the livestock because it is the only one with that characteristic
- Alex/Jessica - key_id can be combined but needs indicator of Tanzania or Mozambique to eliminate duplication. Need to address asterisks so the numbers can be used as data (create separate column for cowshed/barn? Did other barns get added to rooms #?). Normalize terms to be the same (ie. roof_type, remove underscore). Discard data if no way of determining true data (ie. rooms -99)
- Rachel/Jamie - unique identifier needed for interviewers, i.e. m1, m2 for Mozambique; also metadata consistency in the case of "mabati_sloping" versus "mabatisloping"; convert all responses to numerical answers?; questions about the "looks after cows" data (is this a human job); barn question should have its own column; is it better to have more key ids
- Separating out the livestock_owned_and_numbers into two different variables; the plots variable has inconsistent coding; changing the -999 missing data indicator to a black cell or a NA; inconsistencies in roof type answers; removing the asterick in the Tanzania sheet and using another method; livestock table -both numeric and character - needs consistent coding; Look after cows variable - what does that mean?; too much extra space in the headers and columns; remove spaces in variable names and replace with underscore
- Jessea/Nick - reformatting Mozambique livestock table into similar format as Tanzania, reformat the notes to highlighting, fix negative numbers, reformat plots to all y/n
- Danielle/Elaine - normalize values for cells (y/n, #, understores between words, correct mispellings); Place column to combine tabs; ensure there are no blank fields (write 0), number fields should not contain any other characters (comments should be in a seperate column)
- Ashley/Cat: variable names are misspelled, inconsistent data entry, missing data
- Melissa/Michele: consistent values for cells; missing Plots section in Tanzania; etc. - see everyone else's comments. =)
- Erin/Kelly: have a clear interview id system that allows data from both collection sites to be combined into one sheet; allow for an "other" response for when answer is not yes/no.;
- Lorelei/Sanjeet: Agree, maybe use a "notes" field to capture exceptional situations (though you'd have to use qualitative analysis on those fields)
- Are these keyids all really refering to the same farm? Or is this 3 people with 3 clipboards numbering uniquely? <- good question! +1
- Separate Livestock_owned_ and Numbered into four colums. One for each type of livestock. owned (yes .no) and numbers.
- Astra/Tiffany: Add location indicator to key_id, notes column for errors in data (-99 rooms), change barn/cowshed to a yes/no column separate from rooms, fix typos (i.e., errth)
- Gina/Sarah: make variables consistent, add a column for country or modify the key IDs to indicate country, in the Mozambique table split number of total livestock up to match Tanzania, no plot

information for Tanzania so consider getting rid of it, replace blanks with zeros or indicate data is missing
- Denise/Gabriel - we noticed the variables are not consistent immediately
- Denise/June - adding location in new column when combining, need to differentiate the animals in the Mozambigue Livestock table (do we throw it out?), uncertain about whether to keep plot data since not available for both

Exercise - Metadata

Download a clean version of this dataset and open the file with your spreadsheet program. This data has many more variables that were not included in the messy spreadsheet and is formatted according to tidy data principles.

Discuss this data with a partner and make a list of some of the types of metadata that should be recorded about this dataset. It may be helpful to start by asking yourself, “What is not immediately obvious to me about this data? What questions would I need to know the answers to in order to analyze and interpret this data?”

Make note of your answers here

- Where are the villages?
- Where is the roof type, does this matter -- is it covered by another variable?
- where did the chickens etc go :(Oh I guess it's liv_count
- What is meant by "conflict"? - military/terrorism? something else?
- What is a "member_assoc"?
- "No meals" has a numeric value... per day? per week? What is being considered a meal? Does the livestock from messy data have an impact on this? +1
- What is the type of conflict?
- "no-meals?""affect-conflicts?" "member-association?"What most of the columns mean, really
- are uuid keyed to anything?
- What is the research question?
- years liv as in years lived in these villages? Or years lived in the residence/home?
- When were the interviews done (relevant for how to interpret notes like having had a cow in April?)
- What does "affect conflict" mean?
- Definitions/descriptions of the variables
- Do you have ReadMe file that defines your parameters? Yes! Could we see your survey instrument?
- Keys to identify variables,
- How is instanceID generated?+1
- Which items owned are important to ask about?
- Does the type of livestock (cows, chickens, etc.) and who cares for them still matter?
- villages / member associations
- Wall type entries have inconsistencies with spaces/characters; use of "no" or "never" as answers in columns H and I; inconsistencies in month spelling/abbreviations
- There needs to be a readme file or other documentation to explain the variable names. That would

solve a good number of these questions about "what does ___ mean?" +1
- Are "items_owned" any items or only items related to work/livelihood?
- Is "no_meals" per day/week/month?
- are items owned selected from a standard list of pre-chosen items? <- I had the same question, looks like it. +1
    - Is "solar_torch" describing the type of the torch or something else?
- What is the research instrument and method? Questions asked? Structured survey or semistructured?
- what do all the items owned have in common and how does that relate to the research?for example some are vehicles but some are technology
- What, if any, options were avaialble for the items owned question; was that free-text or what?
- Could this have a corresponding definitions or directory table that would answer all of these questions that people are pointing out in this exercise? (creting a metadata tab)
- WHat do the headings stand for? like memb_assoc
- formatting question - why use ";" to seperate vaules in a cell...instead of ","?
- How many people live & work on the farm? How many people does the farm feed?
- How is item type determined
- after seeing the raw data, what is being counted as a room?
- Clarify "affect_conflicts".
- The items owned -contains more than one variable should be separate columns?
- What happened to the plots/water use?

Dates as Data

Exercise: What is the date today?

- 45,325 (this is what Excel counts as the days since this date 1900 **Angie: I'm just recording the answer, I had no clue.** LOL)where do you see this?if you undo a formatted date, it will show you the number it has it stored as, and won't undo to your input. (I've had this happen to my Excel data accidentally sometimes)weird! i usually put my months and days in their own column because date formatting makes me anxious lol
- 2024-02-07 at 12:27pm CT
- 2024-02-07 ISO format :D +1
- 02/07/2024 or 2024-02-07
- 20240207
- 20240207
- 2/7/2024 on my computer :)
- 20240207
- February 7, 2024
- 2/7/2024
- 02/07/2024
- 07/02/2024
- 02/07/2024
- 2/7/24
- 02/07/2024
- Shevat 28, 5784
- 2.7.24

February 7, 2024 +1+1
2024/02/07
2/7/2024
Feb 7
7 February 2024
2024.02.07
7/2/24
02072024

Exercise - formulas for tidy dates

SEPARATING DATES INTO COMPONENTS
Download and open the SAFI_dates.xlsx file. This file contains a subset of the data from the SAFI interviews, including the dates on which the interviews were conducted.

Choose the tab of the spreadsheet that corresponds to the way you format dates in your location (either day first DD_MM_YEAR, or month first MM_DD_YEAR).

Extract the components of the date to new columns. For this we can use the built in Excel functions:

=MONTH()
=DAY()
=YEAR()

Apply each of these formulas to its entire column. Make sure the new column is formatted as a number and not as a date.

Default year exercise

Using the same spreadsheet you used for the previous exercise, add another data point in the interview_date column by typing either 11/17 (if your location uses MM/DD formatting) or 17/11 (if your location uses DD/MM formatting). The Day, Month, and Year columns should populate for this new data point. What year is shown in the Year column?
2024 :D
2024

Quality Control

Exercise - Data validation

Apply a new data validation rule to one of the other *numeric columns* in this data table. Discuss with the person sitting next to you what a reasonable rule would be for the column you've selected. Be sure to create an informative input message.

Apply a new data validation rule to one of the other *categorical columns* in this data table. Discuss with the person sitting next to you what a reasonable rule would be for the column you've selected. Be sure to create an informative input message.

Exporting Data as a csv

Exercise - export your own data

1. Open one of your own spreadsheets (or one of the example excel files)
2. From the top menu select File and Save as.
3. In the Format field, from the list, select Comma Separated Values (.csv).
4. Double check the file name and the location where you want to save it and hit Save.

Install OpenRefine

Setup:
https://datacarpentry.org/openrefine-socialsci/index.html#setup

Download the SAFI dataset:
https://ndownloader.figshare.com/files/11502815

Exercise 1:
Use this same strategy to remove the single quote marks ('), the right square brackets (]), and spaces from the items_owned column.

1. Answer these questions:
    1. What roof types are selected by this procedure?

mabatipitched (10) and mabatisloping (48) + 2+1+1
mabaticpitched 10 and mabatisloping 48
text facet - mabatipitched10
mabatisloping48

    1. How would you restrict this to only one of the roof types?

Add another text facet, then select one of the options
Use the name of one type of roof, or for this one just adding -ip or -is works. Enough to make it only fit one.+1
After creating the facet, click on the one you'd like to edit further
include more letters in the text filter

Day one wrapup one up one down:

Positives

- Intro to OpenRefine! Already know how to apply to some data at work
- OpenRefine transforms save a lot of time
- Great introduction and demo/hands on to cleaning up data using two different tools from both

presenters. It's so challenging to bring so many attendees along in this type of workshop, and everyone was wonderful.

- OpenRefine's facets will allow me to keep using one of my favorite features in our current ILS, even when we move systems. Instant win!
- Open refine seems to help you save time as you are editing data. great step by step process.
- Day, month, and year functions in Excel - didn't know about those functions
- Thanks for reinforcing good data entry rules as a basis for spreadsheet data work!
- Reminder that Spreadsheets should tell a story ... it is should be the same story to each person looking at the spreadsheet. If done well the first time, excel is very powerful.
- I liked the basics of data entry
- Enjoyed getting some step-by-step learning in OpenRefine
- Totally did not know the 1904/1900 year thing. Also, data validation! Excel is one of those programs that everyone just assumes everyone knows how to use...but there are so many little functions in it that you don't automatically know. There was never an excel class in high school, at least for me anyway+2 I've used Excel for years but didn't think to look up Day(), Month() etc.
- Learned so much, Excel, Open Refine. The instructors and supporting documentation was great! I need to practice a bit, but looking forward to doing that and gaining more confidence.
- Learning a lot! And really enjoying the takeaways!
- Liked how deliberate the navigation between different resoruces was provided using the chat links vs etherpad. Plus (more than one) etherpad is awesome. +1 I had never used Etherpad before.+1
- It was much easier to follow than I expected. I learned a lot of useful info about Excel and OpenRefine.
- OpenRefine seems like it could be so helpful and could save so much time! +1
- Thank you IRDL for getting this carpentries workshop set up. Doing the tutorial together with a facilitator and other people in a workshop setting helped me slow down, absorb what I was learning, and learn from other people's questions and different ways of looking at the same spreadsheet.+1

Its good that I can begin using this now on some data that was cleaning up in excel- 'find and replace' feature.
Wish i knew about openrefine earlier! I can see its applicability immediately and can't wait to play with it more

Negatives

- My biggest struggle using OpenRefine is translating what I want to do (plain language) into what OpenRefine options allows for (facet, filter, grel value). More practice in this when showing examples would help.
- I am confused why we would use clustered over facet transforms and a general edit option. Is this just different ways to do the same thing?
- I was a little unclear about the split function, I got lost a little during that part+1
- Agreed about the split function section - I didn't understand what it did and why we were using it+1
- I felt a bit lost at times during the open refine intro
- If you have trouble with the conceptual stage of your project, or change plans midway after you've gathered data, there is not much you can do to "fix" that in the data cleanup/analysis stage.
- Bit of a long day - all very intersting and fun with the instructors - but I should've stretched more :) +1 would have like a few more microbreaks just to clear my head
- Not sure where I would find GREL syntax on my own/if I forget. (my first thought was to ask

ChatGPT...)
- It is difficult to remember the sintax of the replace and split functions and you clearly need to know where the "help' is for these functions.
- I struggled a bit going back and forth between etherpad, Zoom, my own app, and keeping up, but I think the pacing was overall pretty good. I should have just had two screens! +2

Suggest using a simplified dataset for demoing OR to make it easier to follow +1 most of the columns were extraneous (though maybe that is also more realistic - it shows us how filtering and faceting reduces the cognitive overload of all those rows and columns of data)
I don't know what I don't know - my biggest takeaway is that openrefine exists, but I fear that when I'm ready to get my hands dirty with my own data set I'm not going to remember all of the tips Lyrric shared today. (but this is not any fault of the presenters!)
I'm using 2 screens - my cell phone for the presentation and my computer for the "doing" - when you put items in the chat I have a difficult time getting it back to my computer.
There are a lot of tools in OpenRefine. It got to be a bit overwhelming towards the end.
Sometimes it takes me a bit to figure out why we are doing something, but then it all makes sense. But that's a "me" issue.
If you missed a step it was sometimes hard to get back on track, but at least we could watch the screen share.

It took my computer ages to open the OpenRefine file for preview, I wonder if there is a limit to the size of the data set ?

Day 2 Feb 8, 2024

Helpers for today: Jessica Serrao, Sarah Christensen, Courtney Block, Simon Robins
| 9am-2pm Pacific | 10am-3pm Mountain | 11am-4pm Central | 12 noon-5pm Eastern |
Session 1 Open Refine for Social Science**, Part 2: (Lyrric)**

- 0:00 - 0:45 Using Scripts
- **00:45-01:00 Short break**
- 01:00 - 01:30 Exporting and Saving Data from OpenRefine
- 01:30 - 02:00 Other Resources in OpenRefine

Long break: 02:00-3:00

Session 2 Data Analysis and Visualization with R for Social Scientists Part I (Jia Qi Beh)

- 03:00 - 03:45 Before we start (needed?)
- **Short break 3:45 - 4:00**
- 04:00 - 04:30 Introduction to R
- 04:30 - 05:00 Starting with Data

Install R: https://cran.r-project.org/bin/macosx/
Install RStudio: https://posit.co/download/rstudio-desktop/

```
getwd() # tells you your current working directory
# save this code to my local computer - Command+S for Mac ; Ctrl+S for Windows
# run code - Cmd+Return on Mac
```

```
setwd("~/Downloads") # change directory to Downloads
getwd()

dir.create("data") # create a new directory named data
# create two more with the names "data_output" and "fig_output"
dir.create("data_output")
dir.create("fig_output")

# download our input file - the file that we will be working on
download.file(
  "https://raw.githubusercontent.com/datacarpentry/r-socialsci/main/episodes/data/SAFI_clean.csv",
  "data/SAFI_clean.csv", mode = "wb")

# install tidyverse package from CRAN
install.packages("tidyverse")



3 + 4
3 * 4
3 / 4

area_hectares <- 1.0
(area_hectares <- 1.0)
(area_hectares <- 3 + 4)

rm(area_heatares)
area_hectares # print out the value of your object

area_hectares * 2
area_hectares * 10

# find the square root of 10
product_of_sqrt <- sqrt(10)
product_of_sqrt <- sqrt(10)
round(3.14159, digit = 2) # round off to 2 decimal places

?round
?getwd
??squareroot

# Vectors
hh_members <- c(3, 7, 10, 6)
# indexing starts at 1 in R
hh_members[2]
hh_members[2:4]

names <- c("Marie", "Jia", "Lyrric")
names
```

```
names[1]
names[1:3]

num_char <- c(1, 2, 3, "Marie")
tricky <- c(1, 2, 3, 4)

# Conditional subsetting
hh_members > 5
```

Day 3 Feb 9, 2024

Helpers for today: Sarah Christensen, Courtney Block, Simon Robins Fengxia Yan
| 9am-2pm Pacific | 10am-3pm Mountain | 11am-4pm Central | 12 noon-5pm Eastern |
Session 1 **Data Analysis and Visualization with R for Social Scientists Part II** (Lyrric)

- 0:00 - 0:45 Data Wrangling with dplyr
- **00:45-01:00** Short break
- 01:00 - 02:00 Data Wrangling with tidyr

Sign In Here:
Lyrric Jaclkson, ljacks@spelman.edu, she/her, Red Lentil Rounds by Calbee Farms (BBQ)
Angie Rathmel (she/her): University of Kansas, aroads@ku.edu | atruthbrarian@gmail.com | FAV SNACK: Popcorn
Simon Robins (kettle chips, jalepeno)
Jesse Randolph (Pirate's Booty), jrandolph@scu.edu Sarah Smith (trail mix)
Ruth Monnier, she/hers, - rmonnier@pittstate.edu -> dark chocolate, or most sweets,
Maria Waselus, mwaselus@umich.edu, she/her
Catherine Baird,
Wen-ying Lu
Nick D'Andrea, nicholas.dandrea@pepperdin.edu, he/him/his, oreos
Danielle Worthy-Trader Joes unsalted peanut butter filled pretzels
Michele Gibney, University of the Pacific, Yishi Oatmeal (Taro Bubble Tea is so good)
Gina Cherundolo, peanut butter pretzels
Steph Gritz
Kelly Durkin Ruth (she/her), durkinru@usna.edu, twizzlers and pretzel nubs
J. Denice Lewis, lewisjd@wfu.edu, she/her/hers, (favorite snack varies . . . . granola, kettle brand kettle chips, dark chocolate covered caramels with sea salt)
Daniel Keough (he/him), Hoover Institution Library & Archives, dkeough@stanford.edu -- any BBQ flavored chips, trail mix with chocolate
Denise Gehring - nuts and chocolate
Sanjeet Mann (snack <- something with peanut butter)
Lorelei Rutledge (they/them) popcorn!!!! <- plain, butter added, flavor salt? tell us more!
Astra Yatroussis (They/She), Loyola Marymount University: Spicy Tuna Onigiri
Elaine Yeung (she/her), Chapman University, eyeung@chapman.edu (cheez-its)
Chris Marino - Voodoo Chips all the way!+1
Sarah Christensen - pirates booty/nerds gummy clusters

Fengxia Yan(Helper), Morehouse School of Medicine
Jennifer Embree (she/her), Binghamton University, jembree@binghamton.edu (brownies)
Maja Subelj, UA
Ashley Wilson (she/her) Triscuits <- what flavor Triscuits do you like?
Marie, marie.kennedy@lmu.edu - Honey Nut Cheerios (dry, not with milk)# yum!
Sarah Follmer, University of Illinois at Urbana Champaign
Lorna Dawes, of Nebraska
Bradley Dusing, Trinity University, bdusing@trinity.edu, he/she/they, chips and salsa

Catch up on your Code

```
dir.create("data")
dir.create("data_output")
dir.create("fig_output")
download.file(
  "https://raw.githubusercontent.com/datacarpentry/r-socialsci/main/episodes/data/SAFI_clean.csv",
  "data/SAFI_clean.csv", mode = "wb"
)
install.packages("tidyverse")
install.packages("here")
install.packages("libraries")
3 + 5
12 / 7
area_hectares <- 1.0
area_hectares <- 1.0
(area_hectares <- 1.0)
area_hectares
2.47 * area_hectares
area_hectares <- 2.5
2.47 * area_hectares
area_acres <- 2.47 * area_hectares
area_hectares <- 50
area_acres <- 2.47 * area_hectares
area_hectares <- 1.0
area_acres <- area_hectares
area_acres
r_length <- 2.5
r_width <- 3.2
r_area <- r_length * r_width
r_area

# change the values of r_length and r_width
r_length <- 7.0
r_width <- 6.5
# the value of r_area isn't changed
r_area

#Functions and their arguments
a <- 9
b <- sqrt(a)
round(3.14159)
```

```
args(round)
function (x, digits = 0)

  ?round
round(3.14159, digits = 2)
round(3.14159, 2)

round(digits = 2, x = 3.14159)

#Vectors and Data Types
hh_members <- c(3, 7, 10, 6)
hh_members

respondent_wall_type <- c("muddaub", "burntbricks", "sunbricks")
respondent_wall_type

length(hh_members)
length(respondent_wall_type)
typeof(hh_members)
typeof(respondent_wall_type)
str(hh_members)
str(respondent_wall_type)

possessions <- c("bicycle", "radio", "television")
possessions <- c(possessions, "mobile_phone") # add to the end of the vector
possessions <- c("car", possessions) # add to the beginning of the vector
possessions

num_char <- c(1, 2, 3, "a")
num_logical <- c(1, 2, 3, TRUE)
char_logical <- c("a", "b", "c", TRUE)
tricky <- c(1, 2, 3, "4")
num_logical <- c(1, 2, 3, TRUE)
char_logical <- c("a", "b", "c", TRUE)
combined_logical <- c(num_logical, char_logical)

respondent_wall_type <- c("muddaub", "burntbricks", "sunbricks")
respondent_wall_type[2]
respondent_wall_type[c(3, 2)]

more_respondent_wall_type <- respondent_wall_type[c(1, 2, 3, 2, 1, 3)]
more_respondent_wall_type
```

Exercise 1:

1. Using this vector of rooms, create a new vector with the NAs removed.

```
rooms <- c(1, 2, 1, 1, NA, 3, 1, 3, 2, 1, 1, 8, 3, 1, NA, 1)
```

2. Use the function median() to calculate the median of the rooms vector.
3. Use R to figure out how many households in the set use more than 2 rooms for sleeping.

Answers Exercise:

```
# Exercise 1- Part 1
rooms <- c(1, 2, 1, 1, NA, 3, 1, 3, 2, 1, 1, 8, 3, 1, NA, 1)
rooms_no_na <- rooms[!is.na(rooms)]

#Exercise 1- Part 2
median(rooms, na.rm= TRUE)

#Exercise 1- Part 3
rooms_above_2 <- rooms_no_na[rooms_no_na > 2]
length(rooms_above_2)
```

Long break: 02:00-3:00

```
### Data Wrangling with dplyr ###

library(tidyverse)
library(here)

getwd()

interviews <- read_csv("data", "SAFI_clean.csv", na = "NULL")

# inspect the data
interviews
View(interviews)

# select() for selecting columns
selected_cols <- select(interviews, village, no_membrs, respondent_wall_type)
# you can perform the same function using subsetting
interviews[c("village", "no_membrs","respondent_wall_type")]
select(interviews, village:respondent_wall_type)

# filter() to filter rows
filter_God <- filter(interviews, village == "God")
View(filter_God)

filter(interviews, village == "God",
       rooms > 1,
       no_meals > 2)
# instead of using commas, you can also use the & operator
filter(interviews, village == "God" &
       rooms > 1 &
       no_meals > 2)

# use the OR operator (|) to filter rows that meet either one of the conditions
filter_variables <- filter(interviews, village == "God" |
         rooms > 1 |
```

```
        no_meals > 2)
View(filter_variables)

# Pipe operator to combine two or more functions
interviews2 <- filter(interviews, village == "God")
interviews_ch <- select(interviews, village:respondent_wall_type)
```

Exercise:
Using pipes, subset the interviews data to include interviews where respondents were members of an irrigation association (memb_assoc) and retain only the columns affect_conflicts, liv_count, and no_meals.

Session 2 Data Analysis and Visualization with R for Social Scientists Part (Jia Qi Beh)

- 03:00 - 03:50 Data Visualisation with ggplot2
- **3:50 - 4:00 Short break**
- 04:00 - 04:20 Getting started with R Markdown (Optional)
- 04:20 - 04:50 Getting started with Processing JSON data (Optional)
- 04:50 - 05:00 Final wrap up and Post survey
- Link to Post survey https://carpentries.typeform.com/to/UgVdRQ?slug=2024-02-07-lmu-online

```
library(tidyverse)

# run these codes before starting ggplot
## Not run, but can be used to load in data from previous lesson!
interviews_plotting <- interviews %>%
  ## pivot wider by items_owned
  separate_rows(items_owned, sep = ";") %>%
  ## if there were no items listed, changing NA to no_listed_items
  replace_na(list(items_owned = "no_listed_items")) %>%
  mutate(items_owned_logical = TRUE) %>%
  pivot_wider(names_from = items_owned,
              values_from = items_owned_logical,
              values_fill = list(items_owned_logical = FALSE)) %>%
  ## pivot wider by months_lack_food
  separate_rows(months_lack_food, sep = ";") %>%
  mutate(months_lack_food_logical = TRUE) %>%
  pivot_wider(names_from = months_lack_food,
              values_from = months_lack_food_logical,
              values_fill = list(months_lack_food_logical = FALSE)) %>%
  ## add some summary columns
  mutate(number_months_lack_food = rowSums(select(., Jan:May))) %>%
  mutate(number_items = rowSums(select(., bicycle:car)))
```

Colour palette generator: https://coolors.co/

```
# plus "+" operator allows you to add more layers to an existing plot in ggplot

View(interviews_plotting)

# let's make a scatterplot that shows x=no_membrs and y=number_items
interviews_plotting %>%
  ggplot(aes(x = no_membrs, y = number_items)) +
  geom_point()

??scatterplot

# reduce overlapping of data points by setting the transparency
interviews_plotting %>%
  ggplot(aes(x = no_membrs, y = number_items)) +
  geom_point(alpha = 0.3)

# alternatively, reduce overlapping using geom_jitter()
interviews_plotting %>%
  ggplot(aes(x = no_membrs, y = number_items)) +
  geom_jitter()

interviews_plotting %>%
  ggplot(aes(x = no_membrs, y = number_items)) +
  geom_jitter(alpha = 0.3,
              color = "#FCBF49")

# Try making a scatterplot with x=village and y=rooms
interviews_plotting %>%
  ggplot(aes(x = village, y = rooms)) +
  geom_jitter(alpha = 0.3,
              color = "red")

# Make boxplot
interviews_plotting %>%
  ggplot(aes(x = village, y = rooms)) +
  geom_boxplot() +
  geom_jitter(alpha = 0.3,
              color = "red")

# Make violin plot
interviews_plotting %>%
  ggplot(aes(x = village, y = rooms)) +
  geom_violin() +
  geom_jitter(alpha = 0.3,
              color = "red")

# Make barplots
interviews_plotting %>%
  ggplot(aes(x = respondent_wall_type))+
```

```
  geom_bar()

interviews_plotting %>%
  ggplot(aes(x = respondent_wall_type))+
  geom_bar(aes(fill = village)) # to sub-group bars according to village type

interviews_plotting %>%
  ggplot(aes(x = respondent_wall_type))+
  geom_bar(aes(fill = village), position = "dodge")

# add labels to plot
interviews_plotting %>%
  ggplot(aes(x = respondent_wall_type))+
  geom_bar(aes(fill = village), position = "dodge") +
  labs(title = "Proportion of wall type by village",
       x = "Village",
       y = "Percent")

# Faceting
interviews_plotting %>%
  ggplot(aes(x = no_membrs, y = number_items)) +
  geom_jitter()

interviews_plotting %>%
  ggplot(aes(x = no_membrs, y = number_items)) +
  geom_jitter()+
  facet_wrap(~ village)

plot <- interviews_plotting %>%
  ggplot(aes(x = no_membrs, y = number_items)) +
  geom_jitter()+
  facet_wrap(~ village)

ggsave("data_output/plot.pdf", plot, width = 15, height = 10)
```
