Welcome to The Carpentries Etherpad!

This pad is synchronized as you type, so that everyone viewing this page sees the same text. This allows you to collaborate seamlessly on documents.

Use of this service is restricted to members of The Carpentries community; this is not for general purpose use (for that, try https://etherpad.wikimedia.org).

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

---

Course webpage: University of Edinburgh: 3-6 May 2022 (edcarp.github.io)
Pre-workshop survey: The Carpentries Pre-Workshop Survey (typeform.com)
installation instructions:
data download link: https://ndownloader.figshare.com/files/11492171
Data also downloadable using following command in R:
download.file("https://ndownloader.figshare.com/files/11492171","data/SAFI_clean.csv", mode = "wb")

Attendance (write your name):

Day 1: Spreadsheets and OpenRefine

- **10:00-10:10 Intro**
- **10:10 -11:00 Data organisation with Spreadsheets (Andrzej)**
- **11:00-11:05 Coffee break 1**
- **11:05-12:00 Data organisation with Spreadsheets (Andrzej)**
- **12:00-12:45 Lunch break**
- **13:45-14:00 Data cleaning with OpenRefine (Andrzej)**
- **14:00-14:05 Coffee break 2**
- **14:15-15:00 Data cleaning with OpenRefine (Andrzej)**

Before you start (Spreadsheets, https://datacarpentry.org/spreadsheets-socialsci/):

*Make sure you have a spreadsheet program installed before the workshop (**Excel, Apple Numbers** etc.).
A free spreadsheet editor (part of the **Libre Office** open-source project) can be downloaded here:
https://www.libreoffice.org/download/download/
For installation instructions on different operating software, see this page:
https://datacarpentry.org/spreadsheets-socialsci/setup.html

*Download all necessary files:
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

*Make sure you have all the files in the same, accesible folder

Presentation used today can be downloaded here:
https://docs.google.com/presentation/d/1lKEVtGBxJFwxAlTfJsIbFVnKBh_b5UOj/edit?usp=sharing&ouid=104347110123739783728&rtpof=true&sd=true

Before you start (OpenRefine, https://datacarpentry.org/openrefine-socialsci/):

*Download OpenRefine package from: https://openrefine.org/download.html
On Windows, you simply unpack the zipped folder. See installation instructions here:
https://datacarpentry.org/openrefine-socialsci/setup.html

* Download the file we will work on here: https://ndownloader.figshare.com/files/11502815

*Best is to have both OpenRefine and the data file in the same folder

Most common issue:

The most common issue encountered is a lack of administrative privileges. If you plan to access the workshop from your study/workplace computer, firstly make sure you have the relevant software preinstalled or can install it on your own. If not, you will have to ask a relevant person in charge (IT/system administrator) to install and prepare software for you. It should be simpler for home computers, with OpenRefine sometimes requiring to open the exectutable file as administrator (on Windows, can be done by right click -> run as administrator, for one time only, or by navigating to properties -> comparability -> run as admin -> apply, to set it permanently).

Notes on SAFI_Messy (Group: Louise, Manuela, Eilidh)
Issues:

- Typographical errors
- Repeats information
- Merged cells
- Arranged into separate tables, in two tabs.
- Tanzania data from 2017, no year for Mozambique data.
- Inconsistent colour coding.
- Cells record numerical or categorical information and sometimes both.
- What does -99 mean? It is an old convention, which may mean no data was recorded for that question.

Fix:

- Need to bring together into one table.
- Comments should be included in that table - they often qualify the data recorded.
- IDs changed to reflect origin and each column the variable.
- One row per ID.

Notes on SAF (Ath's Group: Anna, Carol, Amy)

-Multiple Information over rows
-missing values,0s, -999
-Multiple tables in one spreadsheet

Notes on SAF (Debbie, Ben, Chris)

- -spelling mistakes in observations
- -99 category?
- Arranged in 3 different tables rather than the data being in seperate tabs
- water use variable is a mix of strings and numbers
- data is not in tidy format

Notes on SAFI_Messy (Group 4)
-issues with multiple tabs
- -99 - What does this mean? is it a null? is it a value?
- variables set up in some sort of hierarchical order rather than individual columns
- multiple tables on the same spreadsheets
- entries not uniform (yes and Yes for example)
- extra information in fields
- inconsistent use of spaces and underscores to separate words (e.g. roof_type or wall type)

Attendance post coffee break (write your name):

Group (Louise, Manuela, Eilidh)

- We need additional information to allow us to understand what we are looking at.
- For example, what does years_liv mean, years lived?
- Where did we get the data from; do the column titles represent questions in the interviews?
- What was allowed as an answer to interview questions?

Attendance post lunch (write your name):

Attendance post afternoon break:

Spreadsheets, presentation from day 1:

Feedback from Attendees on Day 1:

Day 2: Introduction to R

- **10:00-10:15 Intro**
- **10:15 -11:00 R (Elvina)**
- **11:00-11:10 Coffee break 1**
- **11:10-12:00 R (Elvina)**
- **12:00-13:00 Lunch break**
- **13:00-14:00 R (Elvina)**
- **14:00-14:15 Coffee break 2**
- **14:15-15:00 R (Elvina)**

Attendance (write your name):

1. Before we start

Question:
What stats software do you use now?

SAS+
Excel (if that counts! ;))++++++++++
Tableau+
GSheets
Jupyter Notebook
SQL
R,
Graph Pad Prism
Minitab,+
SPSS+
Anaconda, Spyder, Python, Jupyter, STATA

Command to download:

```
download.file("https://ndownloader.figshare.com/files/11492171",
              "data/SAFI_clean.csv", mode = "wb")
```

Area acres?

123.5

Still 6.175 until you re-run the command to multiply it++

Exercise:

Create two variables r_length and r_width and assign them values. It should be noted that, because length is a built-in R function, R Studio might add “()” after you type length and if you leave the parentheses you will get unexpected results. This is why you might see other programmers abbreviate common words. Create a third variable r_area and give it a value based on the current values of r_length and r_width. Show that changing the values of either r_length and r_width does not affect the value of r_area.

Attendance after lunch (write your name):

Lucie Woellenstein (Host)
Athanasia Yiapanas (Helper)
Manuela Mastrogiovanni
Amy Currie
Louise Smith
Cate Lyon
samantha kearsey
Carol Campell
Eilidh MacGlone
Mark Simon Haydn
Chris Cassells
Anna Mazurek
Maram Zahraa (Helper)
Adam Naylor
Debbie McGovernnu

Exercise data types

```
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")
```

Attendance after afternoon break:
Carol Campbell
Athanasia Yiapanas (Helper)
Anna Mazurek
Lucie Woellenstein (Host)
Amy Currie
Louise Smith
Mark Simon Haydn
Samantha Kearsey
Adam Naylor
Cate Lyon
Debbie McGovern
Manuela Mastrogiovanni
Eilidh MacGlone
Maram Zahraa (Helper)
Chris Cassells

3. Starting with Data

```
  # save as r_start_day_01.R in scripts
  # clear workspace
rm(list = ls())
```

```
# to clear console ctl + L

# Load packages we'll be using
library(here)
library(tidyverse)

# Import data  into R
interviews <- read_csv(here("data","SAFI_clean.csv"), na = "NULL")

# Get info about the dataset
interviews
head(interviews) # see first rows
tail(interviews) # see last rows

#dimensions
dim(interviews)
nrow(interviews)
ncol(interviews)

#Summary
str(interviews)
summary(interviewnrs)
glimpse(interviews)

names(interviews)
```

Day 3: Continuation of R

- Extra Content:
- https://www.datacamp.com/ > Datacamp
- https://codeclan.com/ > Codeclan
- https://www.youtube.com/watch?v=MKwyauo8nSI > basics of tidyverse
- https://edcarp.github.io/Ed-DaSH/#:~:text=Ed%2DDaSH%20is%20a%20Data,workshops%20will%20be%20delivered%20remotely. > more Edinburgh carpentries courses
- https://programminghistorian.org/en/lessons/?topic=r
- https://stackoverflow.com/ > StackOverflow: forum to learn anything + everything programming.
- Search for R courses on YouTube > plenty of beginner to advanced tutorials.

Attendance Day 3:
Lucie Woellenstein (host)
Mark Simon Haydn
Carol Campell
Eilidh MacGlone
Amy Currie
Debbie McGovern
Samantha Kearsey

Manuela Mastrogiovanni
Maram Zahraa (Helper)
Anna Mazurek
Cate Lyon
Chris Cassells
Adam Naylor
Louise Smith

Attendance after morning break:
Lucie Woellenstein (Host)
Carol Campbell
Debbie McGovern
Anna Mazurek
Cate Lyon
Louise Smith
Eilidh MacGlone
Amy Currie
Samantha Kearsey
Manuela Mastrogiovanni
Chris Cassells
Maram Zahraa (Helper)
Adam Naylor
Mark Simon Haydn

Attendance after Lunch:
Carol Campbell
Anna Mazurek
Manuela Mastrogiovanni
Louise Smith
Mark Simon Haydn
Cate Lyon
Eilidh MacGlone
Samantha Kearsey
Debbie McGovern
Adam Naylor

Attendance after afternoon break:
Mark Simon Haydn
Samantha Kearsey
Amy Currie
Eilidh MacGlone
Manuela Mastrogiovanni
Carol Campbell
Louise Smith
Anna Mazurek
Debbie McGovern
Cate Lyon
Adam Naylor

Exercise:

1. Create a tibble (interviews_100) containing only the data in row 100 of the interviews dataset.
    1. interviews_100 <- interview[100,]
2. Notice how nrow() gave you the number of rows in the tibble?
    - Use that number to pull out just that last row in the tibble.
    - Compare that with what you see as the last row using tail() to make sure it's meeting expectations.
    - Pull out that last row using nrow() instead of the row number.
    - Create a new tibble (interviews_last) from that last row.
3. Using the number of rows in the interviews dataset that you found in question 2, extract the row that is in the middle of the dataset. Store the content of this middle row in an object named interviews_middle. (hint: This dataset has an odd number of rows, so finding the middle is a bit trickier than dividing n_rows by 2. Use the median( ) function and what you've learned about sequences in R to extract the middle row!
4. Combine nrow() with the - notation above to reproduce the behavior of head(interviews), keeping just the first through 6th rows of the interviews dataset.

Example code:

```
#Q1
# Use interviews[rows,columns] to subset data
# Use interviews[rows,] to just pull out those rows specified
# Assign using = or <- signs
interviews_100 = interviews[100,]

#Q2
# We can nest functions in subset [] brackets
# Pull out last row
interviews[nrow(interviews),]

# tail() pulls out last rows (last 6)
tail(interviews)

# Assign to interviews_last
interviews_last = interviews[nrow(interviews),]


#OR
## 2
# Saving `n_rows` to improve readability and reduce duplication
n_rows <- nrow(interviews)
interviews_last <- interviews[n_rows, ]


#Q3
#Find median row and assign to med_int
```

```
# 1:nrow(interviews) creates a vector containing integers from 1 through nrow(interviews)
# median() finds the median of these
med_int = median(1:nrow(interviews))

#Subset by med_int()
interviews[med_int,]

#or
interviews_middle <- interviews[median(1:n_rows), ]

#Q4
# Use the minus sign to REMOVE these rows.
interviews[-(7:nrow(interviews)),]
```

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.

```
interviews %>% filter(memb_assoc == "yes") %>% select(affect_conflicts, liv_count, no_meals)
```

Exercise

How many households in the survey have an average of two meals per day? Three meals per day? Are there any other numbers of meals represented?

Use group_by() and summari[s|z]e() to find the mean, min, and max number of household members for each village. Also add the number of observations (hint: see ?n).

```
# clean up items_owned - one column per item containing true or false
interviews_items_owned <- interviews %>%
  separate_rows(items_owned, sep = ";") %>%
  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))
```

```
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)))
```

write_csv(interviews_plotting, file = "data_output/interviews_plotting.csv")

Exercise 1:
Use what you just learned to create a scatter plot of rooms by village with the respondent_wall_type showing in different colours. Does this seem like a good way to display the relationship between these variables? What othe