

Data Carpentry Workshop

Leiden, TU Delft, EUR and VU

20th and 21st June 2024

A collaboration of:



Universiteit
Leiden



General information

Code of conduct:

- Use welcoming and inclusive language.
- Be respectful of different viewpoints and experiences.
- Gracefully accept constructive criticism.
- Focus on what is best for the community.
- Show courtesy and respect towards other community members.

→ [The Carpentries Code of Conduct](#)

Instructors and Helpers:

Instructors: Bjørn Bartholdy, Eduard Klapwijk, Peter Verhaar, Stephanie van de Sandt

Helpers: Halford Dace, Elviss Dvinskis, Narmin Rzayeva, Agnes Schneider



Getting help

Stick the red sticky to your laptop screen to indicate you need help to continue; a helper will try to 'unblock' you.

Green sticky notes when you are ready with the exercises

Do you need to catch up with the scripts? A helper will be documenting the scripts in this Google document / GitHub.



Workshop website

<https://ubvu.github.io/2024-06-20-ldev-amsterdam/>

Pre-workshop survey

<https://carpentries.typeform.com/to/wi32rS?slug=2024-06-20-ldev-amsterdam>

Short link to this document: <https://edu.nl/3g7ft>

GS Credits

GS credits TUD

- PhD candidates get 1.5 GS credits if they participate in the two days of the workshop.
- Sign in every day in the 'Roll call', stay for all the sessions, be participative to receive the GS credits.
- You will receive the Course Attendance form with the instructor's signature at the end of the workshop.

GS Credits Leiden

- PhD students will receive 18 hours of study credits.
- Masters students should contact Kristina Hettne - k.m.hettne@library.leidenuniv.nl
- Sign in every day to receive the GS credits.
- You will receive the Course Attendance form with the instructor's signature after the workshop is finished.

GS credits Erasmus University

- [Erasmus Graduate School of Social Sciences and the Humanities \(EGSH\)](#) students can obtain 2.5 ECTS.
- Sign in every day in the 'Roll call', stay for all the sessions, be participative to receive the GS credits.

- Send Eduard (e.klapwijk@essb.eur.nl) a message if you want to receive credits, he will send your name to the EGSH office.

GS credits VU Amsterdam

- Sign in every day to receive 1 EC. We can only issue certificates to those who participate during the whole workshop!
- For PhD candidates: After the workshop, you can request the Data Carpentry edubadge (https://www.edubadges.nl/public/wpj_1EV5QZapbPfDe0GsjA). Follow the instructions to upload the edubadge in Hora Finita.
- Contact Stephanie (s.van.de.sandt@vu.nl) for questions or support.

Short overview of the workshop

 Workshop overview_Data Carpentry for social sciences

Day 2: It is full of R!

Before you leave:

Post-workshop survey

Please help us improve future editions of this workshop by completing the post-workshop survey. Did you learn things you can and will apply in your research or work?

Link to the survey: [Post-workshop Survey](#)

How to continue?

Programme

9:00 - Welcome
9:10 - Starting with data (2)
10:15 - Break
10:30 - Data Wrangling
11:30 - Break
11:45 - Data Wrangling
12:30 - Lunch
13:30 - Quarto
14:25 - Break
14:35 - Data Visualization
15:50 - Break
16:05 - Data Visualization
16:50 - Wrap-up
17:00 - End

List of Participants

- Meron Vermaas, VU
- Moretta Damayanti Fauzi (VU)
- Lisa Schelling (EUR)
- Willem-Jan Gieszen (TUD)
- Mélie Louys (LU)
- Ying-ting Wang (LU)

- Isadora Aibel Olivares (TUD)
- Juliette Lévénez (VU)
- Jorg de Jonge (TUD)
- Muhammad Farooq (VU)
- Hao Zhang (EUR)
- Alberto Estrada (EUR)
- Ibeth Lopez (EUR)
- Iryna Frankova (VU)
- Eszter Szedlacsek (VU)

Day 2 - Questions

You can add questions here and the helpers will be answering them ;-)

-
-
-
-
-

Day 2 Commands and Instructions

Live code:

https://github.com/4TUResearchData-Carpentries/workshop_notes/blob/2406-LDEV-VU/data-carpentry/scripts

Direct link to visualization:

https://github.com/4TUResearchData-Carpentries/workshop_notes/blob/2406-LDEV-VU/data-carpentry/scripts/data-viz.R

PART 1 - Starting with Data in R

Direct link to starting with R code:

https://github.com/4TUResearchData-Carpentries/workshop_notes/blob/2406-LDEV-VU/data-carpentry/scripts/start-data.R

PART 2 - Data wrangling in R

Direct link to Data Wrangling code:

https://github.com/4TUResearchData-Carpentries/workshop_notes/blob/2406-LDEV-VU/data-carpentry/scripts/data-wrangling.R

In case you are interested, you can find [more information about the difference between the base pipe and the magrittr pipe on this blog](#)

<https://datacarpentry.org/r-socialsci/04-tidyr.html#applying-what-we-learned-to-clean-our-data> > link to code block

PART 3 - Introduction to Quarto

<https://cforgaci.github.io/r-socialsci/08-quarto.html>

Direct link to Quarto code:

https://github.com/4TUResearchData-Carpentries/workshop_notes/blob/2406-LDEV-VU/data-carpentry/documents/awesome-report.qmd

https://github.com/4TUResearchData-Carpentries/workshop_notes/tree/2406-LDEV-VU/data-carpentry/documents

PART 4 - Data visualisation in R

Link to crazy visualisations in R from TidyTuesday by 'z3tt':

<https://github.com/z3tt/TidyTuesday>

See also <https://github.com/rfordatascience/tidytuesday>

Direct link to Data viz code:

https://github.com/4TUResearchData-Carpentries/workshop_notes/blob/2406-LDEV-VU/data-carpentry/scripts/data-viz.R

Day 2 Feedback

Please fill in your feedback on the sticky notes: **something you liked** on the yellow note and **something that could be improved** on the orange.

Day 1 - Spreadsheets, OpenRefine, R

List of Participants:

- Jorg de Jonge (TUD)
- Moretta Damayanti Fauzi (VU)
- Mélie Louys (Leiden U)
- Willem-Jan Gieszen (TUD)

- Alberto Estrada (EUR)
- Lisa Schelling (EUR)
- Hao Zhang (EUR)
- Eszter Szedlacsek (VU)
- Muhammad Farooq (VU)
- Juliette Lévénez (VU)
- Anne de Bruijn (VU)
- Isadora Aubel Olivares (TUD)
- Ying-ting Wang (LU)
- Ibeth Lopez (EUR)
- Iryna Frankova (VU)

Programme

9:00 - Welcome
9:15 - Introduction to R
10:00 - Break
10:15 - Intro to R
11:15 - Break
11:30 - Spreadsheets
12:30 - Lunch
13:30 - OpenRefine
14:30 - Break
14:45 - OpenRefine
15:30 - Break
15:45 - Starting with Data
16:50 - Feedback
17:00 - End

For today you need the following files:

- [SAFI_clean.csv](#)
- [SAFI_messy.xlsx](#)
- [SAFI_dates.xlsx](#)
- [SAFI_openrefine.csv](#)

In a folder called [data-carpentry-workshop](#) in an easy to find and access folder, for example within your Home folder. Please include the dashes '-' in between the words!

And the following software:

- R and RStudio:
 - You need to install R before you install RStudio.
 - Instructions:
 - For Windows: <https://datacarpentry.org/r-socialsci/#windows>
 - For MacOS: <https://datacarpentry.org/r-socialsci/#macos>
 - For Linux: <https://datacarpentry.org/r-socialsci/#linux>
 - Please **install the latest release of R studio or update it** if you have previously installed it. You need to have the **version of RStudio (v2023.06)**.
- Spreadsheet programme (i.e. Excel, [LibreOffice](#))
- OpenRefine: Download software from <https://openrefine.org>

Day 1 Questions

You can add questions here and the helpers will be answering them ;-)

-
-
-

Day 1 Commands/Instructions

This section will contain all commands/actions that are done by the instructor. You can go back here if you're stuck or if you missed something.

Slides: [Intro to R](#)

PART 1: Introduction to R

Live code:

https://github.com/4TUResearchData-Carpentries/workshop_notes/tree/2406-LDEV-VU

Direct link to intro to R code:

https://github.com/4TUResearchData-Carpentries/workshop_notes/blob/2406-LDEV-VU/data-carpentry/scripts/intro-to-r.R

Files: for reference, the R lesson uses SAFI_clean.csv. The direct download link for this file is: <https://ndownloader.figshare.com/files/11492171>

Why R

- free and open source
- created by statisticians for statistics
- great for reproducibility
- FREE

We will use RStudio (lots of benefits there as well)

PART 2: Data Organisation in Spreadsheets

Slides:  LDEV_spreadsheets

Exercises:

Working with Messy Data

1. Download the dataset [SAFI_messy.xlsx](#)
2. Open the data in a spreadsheet program
3. Inspect both sheets and compare how the data is structured.
4. 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 analysing the data.
5. Try to harmonise the dataset.
6. Post a screenshot of your result here and add what annoyed you most.

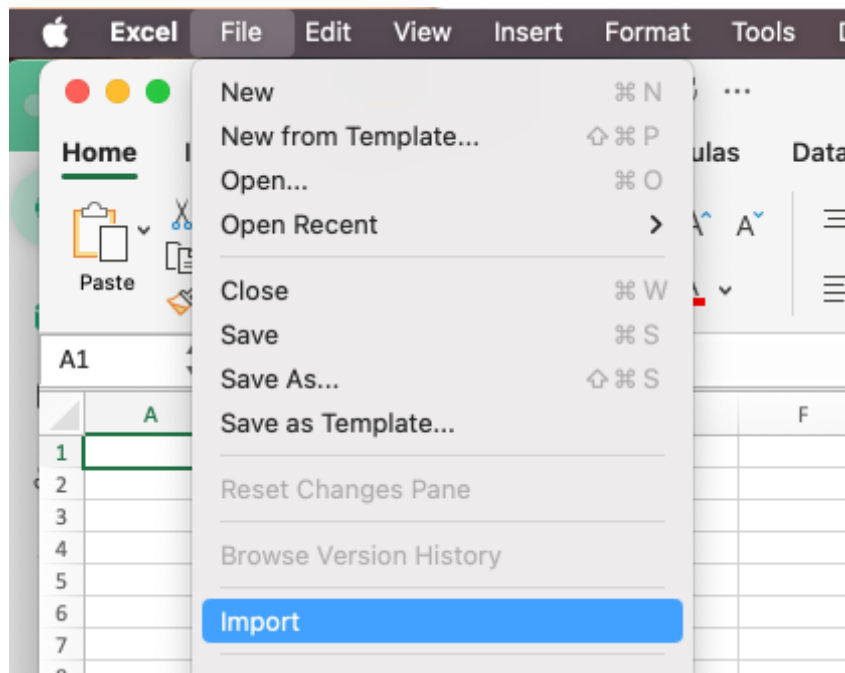
Important: Do not forget our first piece of advice, to create a new file (or tab) for the cleaned data, never modify your original (raw) data.

INSERT SCREEN SHOTS HERE

->
->
->

Importing a .csv

1. Download the clean version of the SAFI dataset: [SAFI_clean.csv](#)
2. Import the comma separated value (csv) file in your spreadsheet program
 - a. Open a new file
 - b. Go to “file” ->



“import”

- c. Select “CSV file” and click “import”
- d. Select the SAFI_clean.csv file from your file system
- e. Press “get data”
- f. Get through the text wizard and select “delimited”
- g. Select “comma” as delimiters
- h. Click “next” and use the “general” column data format
- i. Click “finish”
- j. Keep the “table” selected for “how do you want to view this data?”
- k. Click “Import”

Text Import Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains.

Delimiters

Tab Treat consecutive delimiters as one

Semicolon Text qualifier: "

Comma

Space

Other:

Preview of selected data:

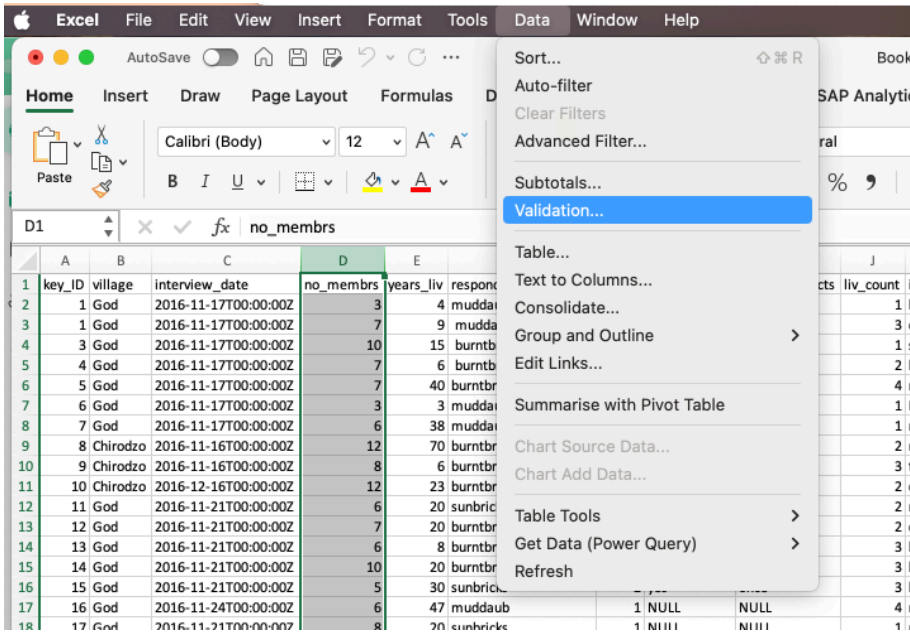
key_ID	village	interview_date	no_membrs	years_liv	respondent_wall_type	rooms	memb_assoc	affect_confl
1	God	2016-11-17T00:00:00Z	3	4	muddaub	1	NULL	NULL
1	God	2016-11-17T00:00:00Z	7	9	muddaub	1	yes	once
3	God	2016-11-17T00:00:00Z	10	15	burntbricks	1	NULL	NULL
4	God	2016-11-17T00:00:00Z	7	6	burntbricks	1	NULL	NULL
5	God	2016-11-17T00:00:00Z	7	40	burntbricks	1	NULL	NULL
6	God	2016-11-17T00:00:00Z	3	3	muddaub	1	NULL	NULL
7	God	2016-11-17T00:00:00Z	6	38	muddaub	1	no	never
8	Chirodzo	2016-11-16T00:00:00Z	12	70	burntbricks	3	yes	never

Exercise: Data Validation

Validate the “no_membrs” column and make sure that only positive values are in there.

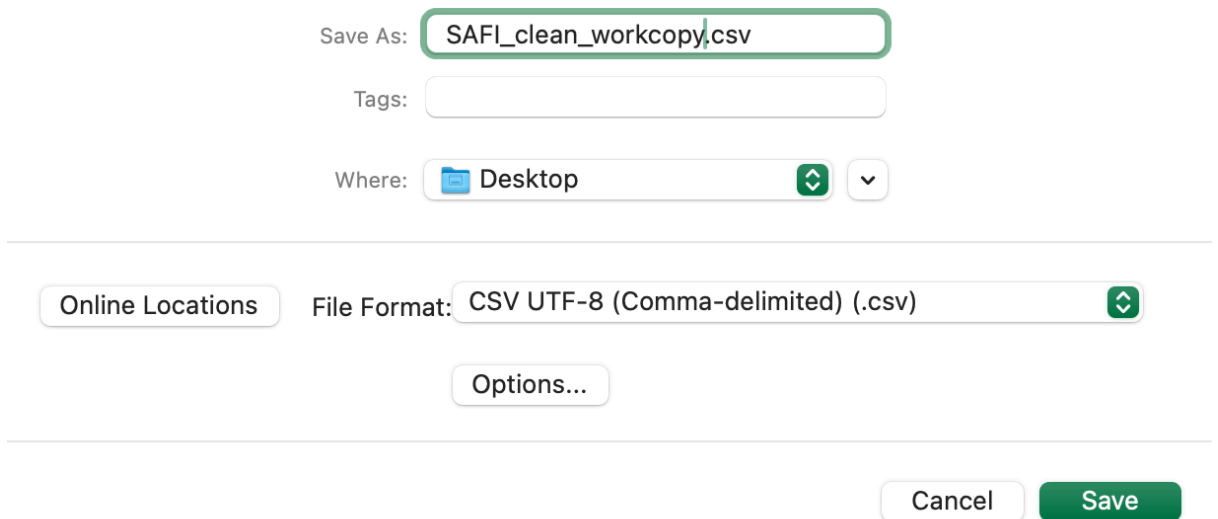
1. Select the column “no_membrs”
2. Click on “Data” -> “Validation”
3. Validation criteria: allow “whole number” ; data: “between”; minimum: “1”, maximum: “30”
4. Click “ok”
5. Try to change a value to 42

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.




Exercise: Exporting Data

1. Click on "File" -> "Save as.."
2. Select csv as file format
3. Click "save"



Notes:

PART 3: OpenRefine

Slides:  2024-06_OpenRefine.pptx

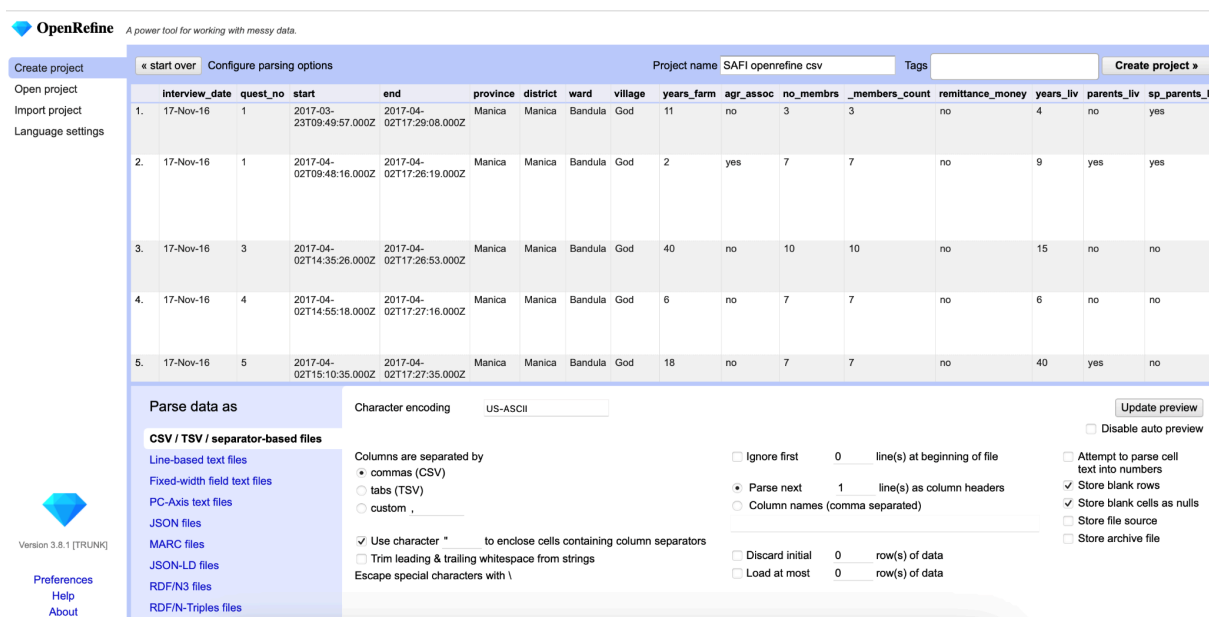
Material: [Lesson material](#)

Data: [SAFI_openrefine.csv](#)

Documentation of instructions

A new project / import data

1. Create Project and select “Get data from This Computer”.
2. Choose the SAFI_openrefine.csv file
3. Click “Next”



The screenshot shows the OpenRefine interface. At the top, the project name is "SAFI openrefine.csv". Below this is a table with 15 columns: interview_date, quest_no, start, end, province, district, ward, village, years_farm, agr_assoc, no_membrs, _members_count, remittance_money, years_liv, parents_liv, and sp_parents_liv. The table contains 5 rows of data. Below the table, the "Parse data as" section is visible, showing options for character encoding (US-ASCII) and parsing options (Commas (CSV), Tabs (TSV), or Custom). The "Parse next" option is selected, and the "Use character" option is checked. The "Update preview" button is visible in the top right corner of the parsing options section.

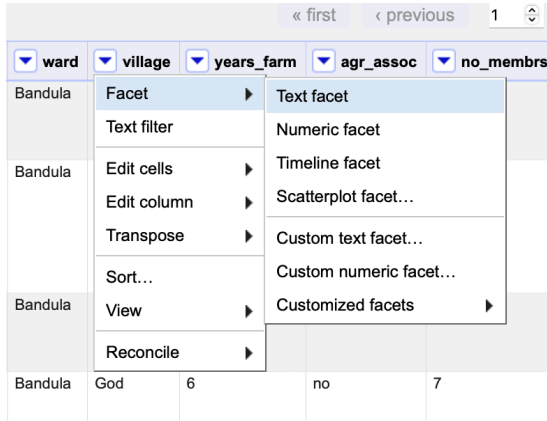
	interview_date	quest_no	start	end	province	district	ward	village	years_farm	agr_assoc	no_membrs	_members_count	remittance_money	years_liv	parents_liv	sp_parents_liv
1.	17-Nov-16	1	2017-03-23T09:49:57.000Z	2017-04-02T17:29:08.000Z	Manica	Manica	Bandula	God	11	no	3	3	no	4	no	yes
2.	17-Nov-16	1	2017-04-02T09:48:16.000Z	2017-04-02T17:26:19.000Z	Manica	Manica	Bandula	God	2	yes	7	7	no	9	yes	yes
3.	17-Nov-16	3	2017-04-02T14:35:26.000Z	2017-04-02T17:26:53.000Z	Manica	Manica	Bandula	God	40	no	10	10	no	15	no	no
4.	17-Nov-16	4	2017-04-02T14:55:18.000Z	2017-04-02T17:27:16.000Z	Manica	Manica	Bandula	God	6	no	7	7	no	6	no	no
5.	17-Nov-16	5	2017-04-02T15:10:35.000Z	2017-04-02T17:27:35.000Z	Manica	Manica	Bandula	God	18	no	7	7	no	40	yes	no

4. Check the preview of the file (s. screenshot)
5. Click “create project”

Using Facets

Text Facets

1. Click on the small arrow next to the variable “village”



- 2.
3. Do you see any problems with the data?
4. Hover the mouse over one of the names in the Facet list. You should see that you have an *edit* function available.

Clustering

1. In the village text facet click on “cluster”
2. Select method -> key collision and key function -> metaphone3

Cluster and edit column "village"

Find groups of different cell values that might be other representations of the same thing. For example, "New York" and "new york" likely refer to the same concept and just differ by capitalization, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method: Key collision Keying function: Metaphone3 Auto-update 2 clusters found

Cluster size	Row count	Values in cluster	Merge?	New cell value
2	45	<ul style="list-style-type: none"> Ruaca (43 rows) Ruca (2 rows) 	<input type="checkbox"/>	Ruaca
2	4	<ul style="list-style-type: none"> Ruaca-Nhamuenda (3 rows) Ruaca - Nhamuenda 	<input type="checkbox"/>	Ruaca-Nhamuenda

Rows in cluster

4 — 45

Average length of choices

4 — 16

Length variance of choices

0.5 — 1

1. Click the Merge? box beside each cluster, then click Merge Selected and Recluster to apply the corrections to the dataset.
2. Close the window

- Clean the misspelt words “Chirdozo” manually with the edit button in the facet window.
- Change Ruaca-Nhamuenda to “Ruaca”

The screenshot shows a data visualization interface with a facet window on the left and a table of records on the right. The facet window is titled 'village' and shows 6 choices: 49, Chirdozo (1), Chirodzo (37), God (43), Ruaca (45), and Ruaca-Nhamuenda (4). An edit dialog box is open for 'Chirdozo'.

id	interview_date	count
1.	2016-11-17T00:00:00Z	1
3.	2016-11-17T00:00:00Z	3
4.	2016-11-17T00:00:00Z	4

Transforming Data

- Click the down arrow at the top of the items_owned column. Choose Edit Cells > Transform...
- First we will remove all of the left square brackets ([]). In the Expression box type `value.replace("[", "")` and click OK.

The screenshot shows a dialog box titled 'Custom text transform on column items_owned'. The Expression box contains the text `value.replace("[", "")`. Below the Expression box is a preview table showing the result of the transformation on five rows of data.

row	value	value.replace("[", "")
1.	['bicycle'; 'television'; 'solar_panel'; 'table']	'bicycle'; 'television'; 'solar_panel'; 'table']
2.	['cow_cart'; 'bicycle'; 'radio'; 'cow_plough'; 'solar_panel'; 'solar_torch'; 'table'; 'mobile_phone']	'cow_cart'; 'bicycle'; 'radio'; 'cow_plough'; 'solar_panel'; 'solar_torch'; 'table'; 'mobile_phone']
3.	['solar_torch']	'solar_torch']
4.	['bicycle'; 'radio'; 'cow_plough'; 'solar_panel'; 'mobile_phone']	'bicycle'; 'radio'; 'cow_plough'; 'solar_panel'; 'mobile_phone']
5.	['motorcyle'; 'radio'; 'cow_plough'; 'mobile_phone']	'motorcyle'; 'radio'; 'cow_plough'; 'mobile_phone']

On error: keep original Re-transform up to 10 times until no change
 set to blank
 store error

OK Cancel

Exercise:

Exercise: [Remove unwanted characters](#) (3 min)

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

Which two items are the most commonly owned? Which are the two least commonly owned?

1. Click the down arrow at the top of the items_owned column. Choose Facet > Custom text facet...
2. In the Expression box, type `value.split(";")`.
3. Sort by count



name	count
mobile_phone	86
radio	86
cow_plough	85
solar_panel	65
bicycle	60
solar_torch	52
table	45
motorcycle	39
television	31
cow_cart	30
stereo	12
NI II L	10

Undo / Redo Changes

1. Click on the Undo / Redo button
2. Restore a version where items_owned was still in square brackets

Filtering

1. Click the down arrow next to respondent_roof_type > Text filter. A respondent_roof_type facet will appear on the left margin.
2. Type mabat

Facet / Filter Undo / Redo 5 / 9 58 matching rows (131 total)

Refresh Reset all Remove all Show as: rows records Show: 5 10 25 50 100 500 1000 rows

respondent_roof_type invert reset

mabat case sensitive regular expression

members_count	remittance_money	years_liv	parents_liv	sp_parents_liv	grand_liv	sp_grand_liv	respondent_roof_type
no		15	no	no	no	no	mabatisloping
no		6	no	no	no	no	mabatisloping
no		70	yes	yes	yes	yes	mabatisloping

Sort

Sort the data by `gps_Altitude`. Do you think the first few entries may have incorrect altitudes?

1. Click on the arrow button next to `gps_Altitude`
2. Click on sort
3. Sort cell values as numbers
4. The option sort has now an option to remove the sort

Numeric Facet

By default, the type of each cell is text. You can change the type by clicking on edit cells, common transforms -> to number.

years_liv	parents_liv	sp_parents_liv	grand_liv	sp_grand_liv	respondent_roof_type
yes		no	yes	grass	m
yes		no	yes	grass	n
6	no				b
40	yes	no	yes		bt
3	no	no	no		m
38	yes	no	yes		m

Facet

Text filter

Edit cells Transform... Common transforms

- Trim leading and trailing whitespace
- Collapse consecutive whitespace
- Unescape HTML entities
- Replace smart quotes with ASCII
- To titlecase
- To uppercase
- To lowercase
- To number
- To date
- To text
- To null
- To empty string

You cannot change a number to "abc" for example

Exercise: Transforming column contents to numbers (3 min)

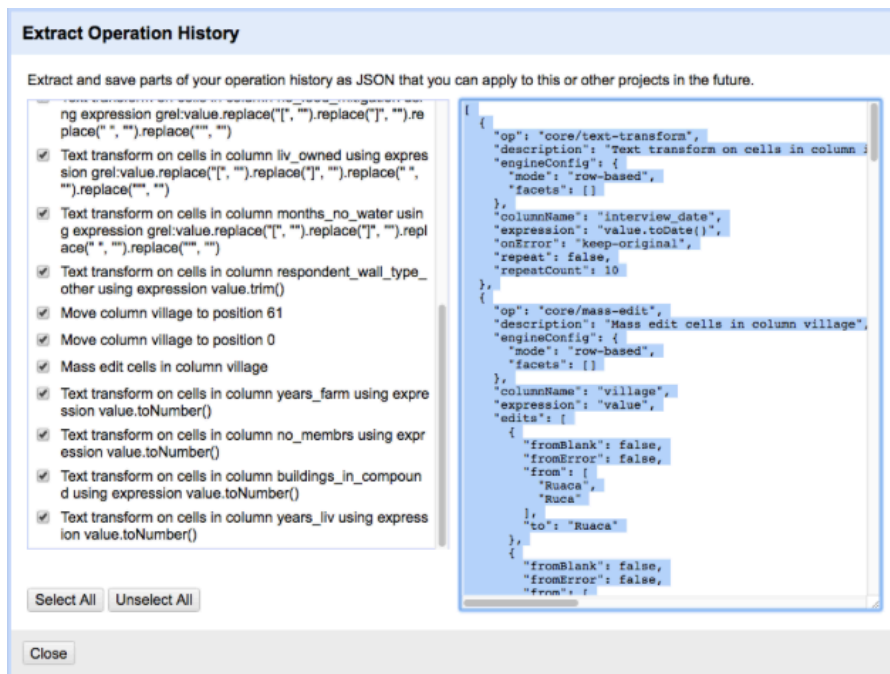
Transform three columns:

- No_membrs
- years_liv
- and buildings_in_compound

from text to numbers. Can all columns be transformed to numbers? - Try it with village for example.

Saving your work as a script

1. In the Undo / Redo section, click Extract..., and select the steps that you want to apply to other datasets by clicking the check boxes.



- 2.

Exporting Cleaned Data

1. Click Export in the top right and select the file type you want to export the data in. Tab-separated values (tsv) or Comma-separated values (csv) would be good choices.

OpenRefine creates a file whose name is based on the project name and asks the browser to download it. Depending on your browser settings, this file is automatically

saved in the default location for downloaded files, or you see a dialog window to choose where you want to save the file.

Useful Links:

OpenRefine has its own web site with documentation and a book:

- [OpenRefine web site](#)
- [OpenRefine User Manual](#)
- [Using OpenRefine](#) book by Ruben Verborgh, Max De Wilde and Aniket Sawant
- [OpenRefine history from Wikipedia](#)

In addition, see these other useful resources:

- [Grateful Data](#) is a fun site with many resources devoted to OpenRefine, including a nice tutorial.
- [OpenRefine source code on GitHub](#)
- Fora: [official forum](#), [StackOverflow](#)

PART 4: Starting with data in R

Live code:

https://github.com/4TUResearchData-Carpentries/workshop_notes/tree/2406-LDEV-VU

- [Lesson material](#) and [slides](#)

Direct link to starting with R code:

https://github.com/4TUResearchData-Carpentries/workshop_notes/blob/2406-LDEV-VU/data-carpentry/scripts/start-data.R

Day 1 Feedback

Please fill in your feedback on the sticky notes: **something you liked** on the green note and **something that could be improved** on the pink.

How to continue with R, research data and research software?

You can join several communities to keep on learning!

- [Leiden Open Science Community](#) / [Research Software Community Leiden](#)
- [Delft Open Science Community](#)
- [Open Science Community Rotterdam](#)
- [EUR Data Coffee Breaks](#)
- [R café at TU Delft](#)
- [R-Ladies Global](#)
- [VU Data Conversation Research Software](#)
- [VU coding café Bytes & Bites](#)

You can join other training and workshops at Leiden, TU Delft, VU and EUR:

- [Workshops/Training at the Centre for Digital Scholarship](#) Leiden
- [Training at TU Delft](#)
- [Training at VU](#)
- [Data Analysis with R](#) at the Erasmus Graduate School for Social Sciences and Humanities
- [Code Refinery](#): intermediate- and advanced level courses on code skills
- [Taxila](#): Dutch national platform that is the one stop shop for training on research data, research software, open science, and the like.

Some online resources for self-study:

- Quarto guidelines: <https://quarto.org/>
- [R for Data Science](#) by Hadley Wickham (online book with exercises)
- [Data Visualization](#) by Kieran Healy (online book with code examples)
- [R Graph Gallery](#) for choosing visualization with tidyverse & ggplot2 (incl. reproducible code)
- [Color in R](#) names and hexcode of colors in R
- [Learning Statistics with R](#) - online book on basics of R and running statistics in R
- [Exploratory Data Analysis in R](#) how to explore your data set
- [R for graduate students](#) Data wrangling and plotting in R
- [R for Non-Programmers: A Guide for Social Scientists](#) resource for beginners
- [Big Book of R](#) “your last-ever bookmark”; a compilation of loads of resources!
- [Applied Statistics with R](#)

Post-workshop survey

Please help us improve future editions of this workshop by completing the post-workshop survey. Did you learn things you can and will apply in your research or work?

Link to the survey: [Post-workshop Survey](#)