

Data Cleaning

Katie Smart, Lars Figenschou
The University Library

25. 09. 2023



Agenda

1. Best practices for organizing data in spreadsheets
 - Break-out exercise: spot the errors
 - Tidy Data Principles
2. Data Cleaning and and Quality Control in Excel
3. Other tools

Handout : [RDM Training @ UiT Zenodo collection](#)

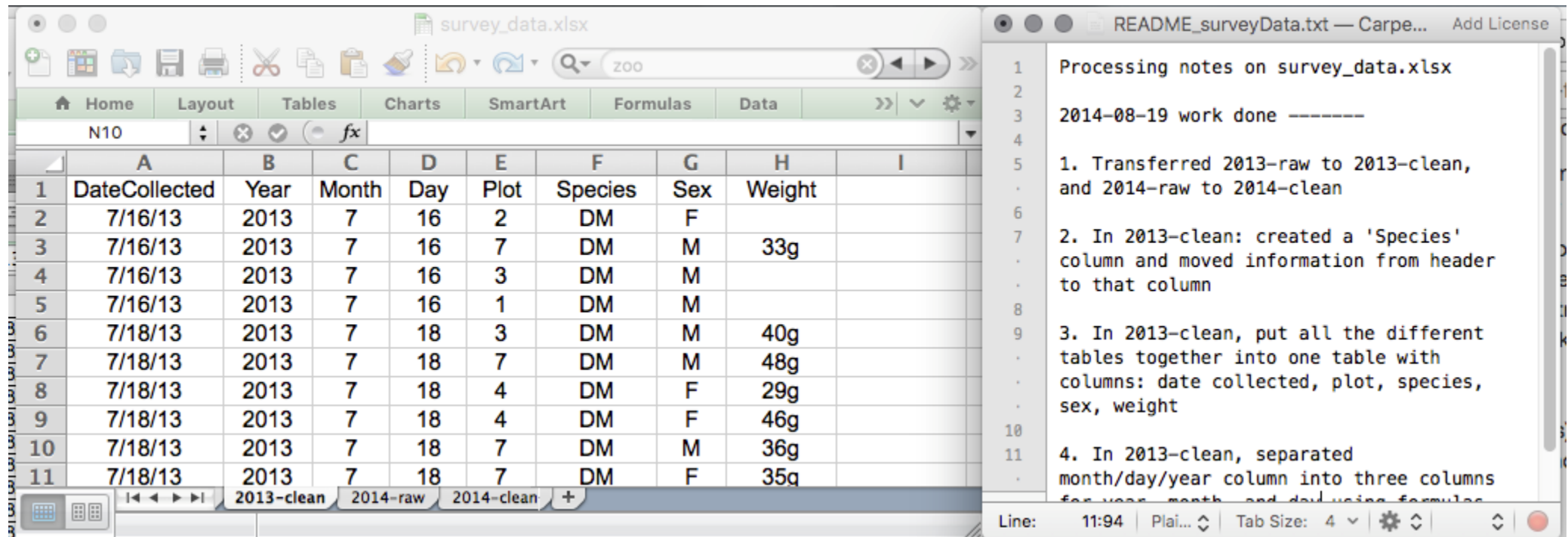
Organizing data in spreadsheets

Goal

- The data should be understood by yourself in the future.
- The data should be understood by others.
(Reusable in FAIR)
- The data should be machine-readable
(Interoperable in FAIR)

Rule of thumb: Never modify your raw data.

- Always make a copy before making changes
- Do not include formulas and calculations into raw-data
- Back up your files
- Keep track of all the steps you take to clean your data in a plain text file (README)



The screenshot shows two windows. The left window is an Excel spreadsheet titled 'survey_data.xlsx' with a 'zoo' search bar. The spreadsheet has columns A through I. The data is as follows:

	A	B	C	D	E	F	G	H	I
1	DateCollected	Year	Month	Day	Plot	Species	Sex	Weight	
2	7/16/13	2013	7	16	2	DM	F		
3	7/16/13	2013	7	16	7	DM	M	33g	
4	7/16/13	2013	7	16	3	DM	M		
5	7/16/13	2013	7	16	1	DM	M		
6	7/18/13	2013	7	18	3	DM	M	40g	
7	7/18/13	2013	7	18	7	DM	M	48g	
8	7/18/13	2013	7	18	4	DM	F	29g	
9	7/18/13	2013	7	18	4	DM	F	46g	
10	7/18/13	2013	7	18	7	DM	M	36g	
11	7/18/13	2013	7	18	7	DM	F	35g	

The right window is a text editor titled 'README_surveyData.txt' with the following content:

```
1 Processing notes on survey_data.xlsx
2
3 2014-08-19 work done -----
4
5 1. Transferred 2013-raw to 2013-clean,
6 and 2014-raw to 2014-clean
7
8 2. In 2013-clean: created a 'Species'
9 column and moved information from header
10 to that column
11
12 3. In 2013-clean, put all the different
13 tables together into one table with
14 columns: date collected, plot, species,
15 sex, weight
16
17 4. In 2013-clean, separated
18 month/day/year column into three columns
19 for year, month, and day using formulas
```

Source: [Data carpentry](#)



Describe your data in a README file

Everything necessary for your future you and others to understand what is in the dataset and be able to reuse it.

- Describe the methods for data collection and processing
- Keep track of all changes
- Variables in columns + unit of measure
- Abbreviations used
- Store it close to the data it describes.

Explanation of column headings used in file Icecream_sales_2020
Column A contains temperature measurements in degrees Celsius
Column B contains date in YYYY-MM-DD
Column C contains money earned per day in NOK (Norwegian kroner)
Column D contains ...

Practical exercise – break out room

Link to spreadsheet in chat: [survey_data_spreadsheet_messy.xls](#)

Discuss:

- What is wrong with the spreadsheet?
- How many mistakes can you find?
- How could the spreadsheet be improved?

5 minutes

The Tidy Data Principles

1. Every variable must have a separate column.
2. Every observation must have a separate row.
3. Only one data point per cell.

country	year	cases	population
Afghanistan	1999	75	1999071
Afghanistan	2000	66	20009360
Brazil	1999	737	17206362
Brazil	2000	488	17404898
China	1999	258	127215272
China	2000	76	128028583

variables

country	year	cases	population
Afghanistan	1999	75	1999071
Afghanistan	2000	66	20009360
Brazil	1999	737	17206362
Brazil	2000	488	17404898
China	1999	258	127215272
China	2000	76	128028583


observations

country	year	cases	population
Afghanistan	99	75	1999071
Afghanistan	00	66	20009360
Brazil	99	737	17206362
Brazil	00	488	17404898
China	99	258	127215272
China	00	76	128028583

values

Only one data point per cell:

Date collected	Plot	Species-Sex	Weight
1/9/78	1	DM-M	40
1/9/78	1	DM-F	36
1/9/78	1	DS-F	135
1/20/78	1	DM-F	39
1/20/78	2	DM-M	43
1/20/78	2	DS-F	144
3/13/78	2	DM-F	51
3/13/78	2	DM-F	44
3/13/78	2	DS-F	146



Date collected	Plot	Species	Sex	Weight
1/9/78	1	DM	M	40
1/9/78	1	DM	F	36
1/9/78	1	DS	F	135
1/20/78	1	DM	F	39
1/20/78	2	DM	M	43
1/20/78	2	DS	F	144
3/13/78	2	DM	F	51
3/13/78	2	DM	F	44
3/13/78	2	DS	F	146

Solution:
Add more columns

Avoid comments and units in the cells

18.07.2013	3 M	40g
18.07.2013	7 M	48g
18.07.2013	4 F	29g
18.07.2013	4 F	46g
18.07.2013	7 M	36g
18.07.2013	7 F	25

Solution:

Add units to the column title or into a separate column.

13.11.2013	17 F	118
13.11.2013	11 F	126
13.11.2013	17 M	132 (scale not calibrated)
13.11.2013	14 F	113 (scale not callibrated)
13.11.2013	11 F	122
13.11.2013	4 F	107
13.11.2013	4 F	115

Solution:

Add the information to a separate column


Source: [Data carpentry](#)



Alternative: Add metadata in the README

No formatting

Plot: 2			
Date collect	Species	Sex	Weight
1/8/14	NA		
1/8/14	DM	M	44
1/8/14	DM	M	38
1/8/14	OL		
1/8/14	PE	M	22
1/8/14	DM	M	38
1/8/14	DM	M	48
1/8/14	DM	M	43
1/8/14	DM	F	35
1/8/14	DM	M	43
1/8/14	DM	F	37
1/8/14	PF	F	7
1/8/14	DM	M	45
1/8/14	OT		
1/8/14	DS	M	157
1/8/14	OX		
2/18/14	NA	M	218
2/18/14	PF	F	7
2/18/14	DM	M	52
	measurement device not calibrated		



Date collect	Species	Sex	Weight	Calibrated
1/8/14	NA			
1/8/14	DM	M	44	Y
1/8/14	DM	M	38	Y
1/8/14	OL			
1/8/14	PE	M	22	Y
1/8/14	DM	M	38	Y
1/8/14	DM	M	48	Y
1/8/14	DM	M	43	Y
1/8/14	DM	F	35	Y
1/8/14	DM	M	43	Y
1/8/14	DM	F	37	Y
1/8/14	PF	F	7	Y
1/8/14	DM	M	45	Y
1/8/14	OT			
1/8/14	DS	M	157	N
1/8/14	OX			
2/18/14	NA	M	218	N
2/18/14	PF	F	7	Y
2/18/14	DM	M	52	Y

Date Collect	Plot	Sex	Weight
19.08.2013	8	F	52
17.10.2013	3	F	33

Do not merge cells

Solution:
Add the information into a new column.

Avoid adding multiple tabs and tables

2013 Field Season

Species: DM				Species: DO				Species: DS			
Date Collect	Plot	Sex	Weight	Date Collect	Plot	Sex	Weight	Date Collec	Plot	Sex	Weight
16.07.2013	2	F		19.08.2013	8	F	52	12.11.2013	9	F	117
16.07.2013	7	M	33g	17.10.2013	3	F	33	12.11.2013	1	F	121
16.07.2013	3	M		17.10.2013	3	F	50	12.11.2013	20	M	115
16.07.2013	1	M		17.10.2013	17	F	48	12.11.2013	9	F	120
18.07.2013	3	M	40g	17.10.2013	17	F	31	13.11.2013	17	F	118
18.07.2013	7	M	48g	18.10.2013	8	F	41	13.11.2013	11	F	126
18.07.2013	4	F	29g	12.11.2013	1	F	44	13.11.2013	17	M	132 (scale not calibrated)
18.07.2013	4	F	46g	12.11.2013	1	M	48	13.11.2013	14	F	113 (scale not callibrated)
18.07.2013	7	M	36g	14.11.2013	8	F	39	13.11.2013	11	F	122
18.07.2013	7	F	35g	10.12.2013	9	F	40	13.11.2013	4	F	107
18.07.2013	8	F	22g	10.12.2013	1	M	45	13.11.2013	4	F	115
18.07.2013	7	F	42g	11.12.2013	8	F	41				
18.07.2013	4	F	41g								
18.07.2013	6	F	37g								

2013 | 2014 | dates | +

Each spreadsheet should only contain one table with data.

Solution:

If possible, combine everything into one table, or store each table as separate files.

Source: [Data carpentry](#)



Column headers

2013 Field Season			
Species: DM			
Date Collect	Plot	Sex	Weight
16.07.2013	2	F	33g
16.07.2013	7	M	33g
16.07.2013	3	M	33g
16.07.2013	1	M	33g
18.07.2013	3	M	40g

- Only one column title
- Avoid spaces and special characters

Column headers

	Habitat		
Species	X	Y	Z
A	0	3	0
B	1	0	2



Species	HabitatX	HabitatY	HabitatZ
A	0	3	0
B	1	0	2



Species	Habitat	Abundance
A	Y	3
B	X	1
B	Z	2



Use short and descriptive column titles

- Avoid special characters:

/ \ : * . ? ' < > [] () & \$ æ Æ ø Ø å Å ä Ä

- Avoid using spaces
 - instead use underscore or CamelCase
- Include units (weight_g)

Good Name	Good Alternative	Avoid
Max_temp_C	MaxTemp	Maximum Temp (°C)
Precipitation_mm	Precipitation	precmm
Mean_year_growth	MeanYearGrowth	Mean growth/year
sex	sex	M/F
weight	weight	w.
cell_type	CellType	Cell Type
Observation_01	first_observation	1st Obs

Use consistent date format

Plot: 2			
Date collected	Species	Sex	Weight
1/8/14	NA		
1/8/14	DM	M	44
1/8/14	DM	M	38
1/8/14	OT		

18.07.2013	3	M	40g
18.07.2013	7	M	48g
18.07.2013	4	F	29g
18.07.2013	4	F	46g
		M	36g
		F	25g

Plot: 3			
Date collected	Species	Sex	Weight
1/8	PF	M	7
2/18	OT	M	24
2/19	OT	F	23
3/11	NA	M	232
3/11	OT	F	22
2/11	OT	M	26

Recommended: Use the international standard format YYYY-MM-DD

Scientists rename human genes to stop Microsoft Excel from misreading them as dates

Sometimes it's easier to rewrite genetics than update Excel

By James Vincent | Aug 6, 2020, 8:44am EDT

f t  SHARE



Illustration by Alex Castro / The Verge

There are tens of thousands of genes in the human genome: minuscule twists of DNA and RNA that combine to express all of the traits and characteristics that make each of us unique. Each gene is given a name and alphanumeric code, known as a symbol, which scientists use to coordinate research. But over the past year or so, some 27 human genes have been renamed, all because Microsoft Excel kept misreading their symbols as dates.


**Verge
deals**

Subscribe to get the best Verge approved tech deals of the week

Email (required)

By signing up, you agree to our [Privacy Policy](#) and European users agree to the [data policy](#).

SUBSCRIBE

Pay attention to the standard format in Excel

Excel date formats can be problematic

	A	B	C
1	DATE	Number	How it was interpreted
2	Jul-10	40360	1-Jul-10
3	Jul-14	41821	1-Jul-14
4	Jul-15	42186	1-Jul-15
5	Jul-22	44743	1-Jul-22

Excel display dates in many different formats, but stores dates as numbers.

Alternatives:

- Store dates as a string YYYYMMDD
- Handle dates as several data points (separate columns for year, month, day)

Data cleaning

- Deleting redundant data
- Separate or combine values
- Conversions
- Grammatical errors
- Inconsistent naming
- Date formats
- Problematic «NULL» -values

Delete redundant data:

Remove
unwanted entries

- Duplicates
- Irrelevant observations
- Incomplete data
- Invalid data
- Conflicting data

Consider carefully whether an observation should be removed!

Use sorting to identify deviating and missing values

- Deviating values will sort to the top or bottom.
- Work through the spreadsheet by sorting for each column and check for invalid entries.

The screenshot displays an Excel spreadsheet with the following data:

F	G	H	I	J
Unnamed: 5	TYPE	CITY (kWh/100 km)	HWY (kWh/100 km)	COMB (kWh/100 km)
30	A1	B	16.4	18.4
30	A1	B	17	18.6
30	A1	B	16.4	18.4
30	A1	B	16.4	18.4
30	A1	B	19	21.1
30	A1	B	19	21.1
19	A1	B	16.4	18.7
19	A1	B	16.4	18.7
19	A1	B	16.4	18.7
19	A1	B	16.4	18.7
38	A1	B	23	23.5
38	A1	B	23	22.5
38	A1	B	22	22.1
15	A1	B	23	22.5
36	A1	B	20	20.7
36	A1	B	23	22.7
36	A1	B	22	21
36	A1	B	20	20.3
10	A1	B	23	23.6
10	A1	B	23	23.6
33	A1	B	22	21.9
33	A1	B	23.8	23.6
33	A1	B	23.2	23.6

The 'Sort & Filter' dropdown menu is open, showing options: Sort A to Z, Sort Z to A, Sort by Color, Sheet View, Clear Filter From "HWY (kWh/100 km)", Filter by Color, Text Filters, and Search. The 'Custom Sort' dialog box is also open, showing a list of values to be sorted: (Select All), 18.8, 19.6, 19.7, 19.8, 20.6, 20.7, 20.8, and 21.

Use conditional formatting to highlight deviations and irregularities

The screenshot displays the Microsoft Excel interface with the 'Conditional Formatting' menu open. The 'Greater Than...' option is highlighted. The background shows a table with the following data:

YEAR	Make	Model	Size	(kW)	Unnamed: 5	TYPE	CITY (kWh/100 km)	HW
2016	NISSAN	LEAF	MID-SIZE	80	A1	B	16.5	20.8
2016	NISSAN	LEAF	MID-SIZE	80	A1	B	17	20.7
2015	NISSAN	LEAF	MID-SIZE	80	A1	B	16.5	20.8
2014	NISSAN	LEAF	MID-SIZE	80	A1	B	16.5	20.8
2013	NISSAN	LEAF	MID-SIZE	80	A1	B	19.3	23
2012	NISSAN	LEAF	MID-SIZE	80	A1	B	19.3	23
2016	MITSUBISHI	i-MiEV	SUBCOMPACT	49	A1	B	16.9	21.4
2015	MITSUBISHI	i-MiEV	SUBCOMPACT	49	A1	B	16.9	21.4
2014	MITSUBISHI	i-MiEV	SUBCOMPACT	49	A1	B	16.9	21.4
2013	MITSUBISHI	i-MiEV	SUBCOMPACT	49	A1	B	16.9	21.4
2012	MITSUBISHI	i-MiEV	SUBCOMPACT	49	A1	B	16.9	21.4
2016	TESLA	MODEL X P90D	SUV - STANDARD	568	A1	B	23.6	23.3
2016	TESLA	MODEL S P85D/P90D	FULL-SIZE	568	A1	B	23.4	21.5
2016	TESLA	MODEL S P90D (R)	FULL-SIZE	568	A1	B	22.9	21
2015	TESLA	MODEL S P85D/P90D	FULL-SIZE	515	A1	B	22.4	21.5

Removing duplicates

The screenshot shows the Microsoft Excel interface with the 'Table Design' tab selected. The 'Remove Duplicates' button is highlighted in the 'Tools' group. A dialog box titled 'Remove Duplicates' is open, displaying a list of columns with checkboxes. The 'My data has headers' checkbox is checked. The columns listed are YEAR, Make, Model3, Size, (kW), and Unnamed: 5, all of which are checked. The background shows a table with columns for YEAR, Make, Model3, Size, (kW), Unnamed: 5, and TYPE. The data includes entries for Nissan Leaf and Mitsubishi i-MiEV.

YEAR	Make	Model3	Size	(kW)	Unnamed: 5	TYPE
2016	NISSAN	LEAF	MID-SIZE	80	A1	B
2016	NISSAN	LEAF	MID-SIZE	80	A1	B
2015	NISSAN	LEAF				B
2014	NISSAN	LEAF				B
2013	NISSAN	LEAF				B
2012	NISSAN	LEAF				B
2016	MITSUBISHI	i-MiEV				B
2015	MITSUBISHI	i-MiEV				B
2014	MITSUBISHI	i-MiEV				B
2013	MITSUBISHI	i-MiEV				B
2012	MITSUBISHI	i-MiEV				B
2016	TESLA	MODEL S 70D	FULL-SIZE	386	A1	B
2016	TESLA	MODEL X 90D	SUV - STANDARD	286	A1	B

Before removing duplicates:
Filter for unique values to confirm that you will get the result you expect.

-> Sort and filterer -> Advanced
-> Unique records only

Problematic null values

Date collect	Species	Sex	Weight	Calibrated
1/8/14	NA			
1/8/14	DM	M	44	Y
1/8/14	DM	M	38	Y
1/8/14	OL			
1/8/14	PE	M	22	Y
1/8/14	DM	M	38	Y
1/8/14	DM	M	48	Y
1/8/14	DM	M	43	Y
1/8/14	DM	F	35	Y
1/8/14	DM	M	43	Y
1/8/14	DM	F	37	Y
1/8/14	PF	F	7	Y
1/8/14	DM	M	45	Y
1/8/14	OT			
1/8/14	DS	M	157	N
1/8/14	OX			
2/18/14	NA	M	218	N
2/18/14	PF	F	7	Y
2/18/14	DM	M	52	Y

Table 1. Commonly used null values, limitations, compatibility with common software and a recommendation regarding whether or not it is a good option. Null values are indicated as compatible with specific software if they work consistently and correctly with that software. For example, the null value "NULL" works correctly for certain applications in R, but does not work in others, so it is not presented in the table as R compatible.

Null values	Problems	Compatibility	Recommendation
0	Indistinguishable from a true zero		Never use
Blank	Hard to distinguish values that are missing from those overlooked on entry. Hard to distinguish blanks from spaces, which behave differently.	R, Python, SQL	Best option
-999, 999	Not recognized as null by many programs without user input. Can be inadvertently entered into calculations.		Avoid
NA, na	Can also be an abbreviation (e.g., North America), can cause problems with data type (turn a numerical column into a text column). NA is more commonly recognized than na.	R	Good option
N/A	An alternate form of NA, but often not compatible with software		Avoid
NULL	Can cause problems with data type	SQL	Good option
None	Uncommon. Can cause problems with data type	Python	Avoid
No data	Uncommon. Can cause problems with data type, contains a space		Avoid
Missing	Uncommon. Can cause problems with data type		Avoid
-,+,,	Uncommon. Can cause problems with data type		Avoid

(White et al., 2013)

Use consistent notation for null-values and document this in the README file!



*Separate or
combine values:
Split text into
several columns*

Convert Text to Columns Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

Delimiters

- Tab
- Semicolon
- Comma
- Space
- Other:

Treat consecutive delimiters as one

Text qualifier:

Data preview

Model	Size	(kW)	Unamed: 5	TYPE	CITY (kWh/100 km)	HWY
NISSAN LEAF	MID-SIZE					
NISSAN LEAF	MID-SIZE					
NISSAN LEAF	MID-SIZE					
NISSAN LEAF	MID-SIZE					
NISSAN LEAF	MID-SIZE					
NISSAN LEAF	MID-SIZE					
NISSAN LEAF	MID-SIZE					
MITSUBISHI i-MiEV	SUBCOMPACT					
MITSUBISHI i-MiEV	SUBCOMPACT					
MITSUBISHI i-MiEV	SUBCOMPACT					
MITSUBISHI i-MiEV	SUBCOMPACT					
MITSUBISHI i-MiEV	SUBCOMPACT					
MITSUBISHI i-MiEV	SUBCOMPACT					
TESLA MODEL X P90D	SUBCOMPACT					
TESLA MODEL S P85D/P90D	FULL-SIZE					
TESLA MODEL S P90D (Refresh)	FULL-SIZE					
TESLA MODEL S P85D/P90D	FULL-SIZE					
TESLA MODEL S 70D	FULL-SIZE					
TESLA MODEL X 90D	SUBCOMPACT					
TESLA MODEL S 85D/90D	FULL-SIZE					
TESLA MODEL S 90D (Refresh)	FULL-SIZE					
TESLA MODEL S PERFORMANCE	FULL-SIZE					
TESLA MODEL S PERFORMANCE	FULL-SIZE					
TESLA MODEL S (60 kWh battery)	FULL-SIZE	283	A1	B	22.2	21.7



2016	NISSAN	LEAF	MID-SIZE
2016	NISSAN	LEAF	MID-SIZE
2015	NISSAN	LEAF	MID-SIZE
2014	NISSAN	LEAF	MID-SIZE
2013	NISSAN	LEAF	MID-SIZE
2012	NISSAN	LEAF	MID-SIZE
2016	MITSUBISHI	i-MiEV	SUBCOMPACT
2015	MITSUBISHI	i-MiEV	SUBCOMPACT
2014	MITSUBISHI	i-MiEV	SUBCOMPACT
2013	MITSUBISHI	i-MiEV	SUBCOMPACT
2012	MITSUBISHI	i-MiEV	SUBCOMPACT

Fix grammatical errors and inconsistencies

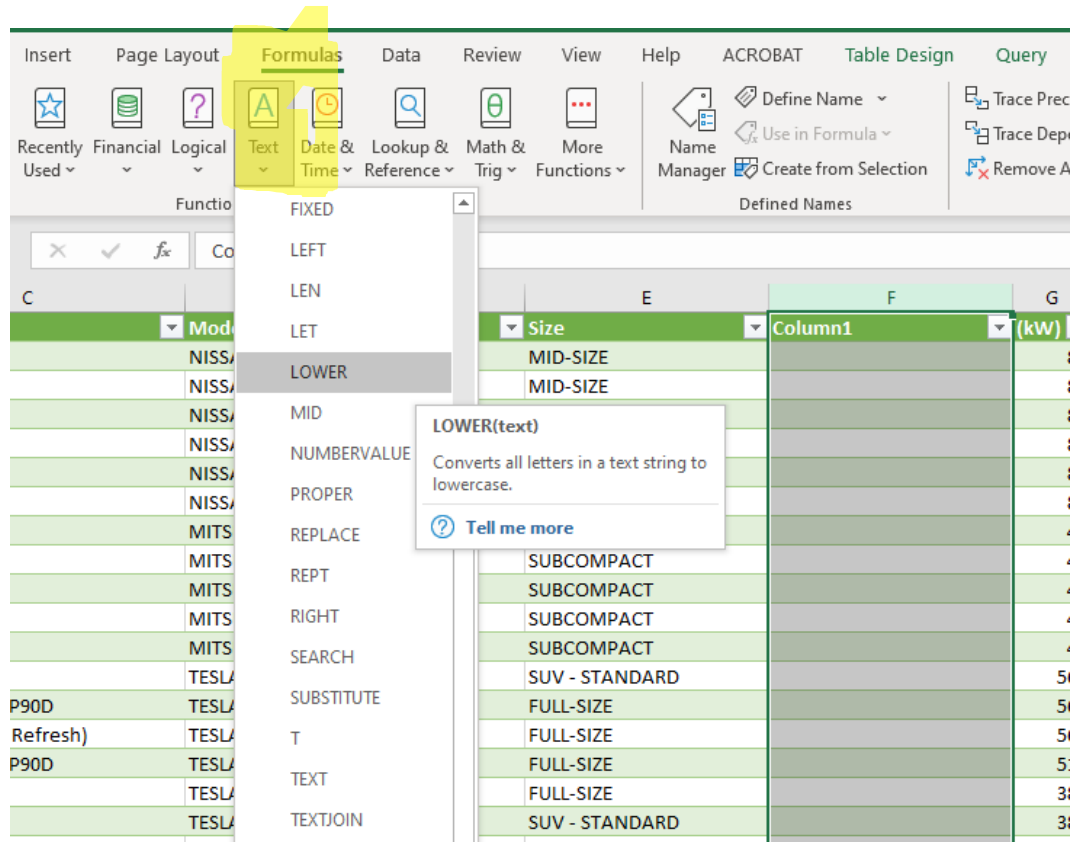
- Grammatical errors
- Inconsistent use of upper and lower case
- Inconsistent titles for columns. Use standardised names across datasets.
- Check for overlapping categories of variables or values. Perhaps they can be combined.

1. Perform tasks that do not require column editing first (spell check or «Find and replace»)
2. Tasks that require column editing.

Steps to change a column:

1. Add a new column (B) next to the original column (A)
2. Add a formula which transforms the data of column (A) to the new column (B).
3. Copy the new column (B), and paste it, **as values**, into the new column (B) – this *removes* the formula
4. Remove the original column (A), and column (B) will be converted to (A).

Use formulas to change text



- Lower – Changes all letters to lower case
- Upper – Changes all letters to upper case
- TRIM – Removes redundant spaces (leading and trailing)
- REPLACE - replaces part of a text string, based on the number of characters you specify, with a different text string
- SUBSTITUTE - substitutes new_text for old_text in a text string

Data validation tools help you avoid entering wrong values

AutoSave ON | survey_data_spreadsheet_messy.xlsx - Compatibility Mode - Saved

Home Insert Draw Page Layout Formulas **Data** Review View Automate Tell me

Get Data (Power Query) From Picture Refresh All Properties Edit Links

Sort Filter Advanced Text to Columns Flash-fill Remove Duplicates **Data Validation** Consolidate What-if Analysis Group Ungroup Subtotal Hide D

B8 | x ✓ fx

2013 Field Season

Species: DS	Date Collecte	Plot	Sex	Weight
	12/11/2013	9	F	117
	12/11/2013	1	F	121
	12/11/2013	20	M	115
	12/11/2013	9	F	120
	13/11/2013	17	F	118
	13/11/2013	11	F	126
	13/11/2013	17	M	132 (scale not calibrated)
	13/11/2013	14	F	113 (scale not callibrated)
	13/11/2013	11	F	122
	13/11/2013	4	F	107
	13/11/2013	4	F	115

x

Data Validation

Settings Input Message Error Alert

Validation criteria

Allow: Whole number Ignore blank

Data: between

Minimum: 1

Maximum: 100

Apply these changes to all other cells with the same settings

Clear All Cancel OK

Prepare the
data for sharing
and archiving

- Remove formatting
- Export the spreadsheets to open and persistent file formats

Remove all formatting

- Also conditional formatting

Clear Formats
Clear only the formatting that is applied to the selected cells.

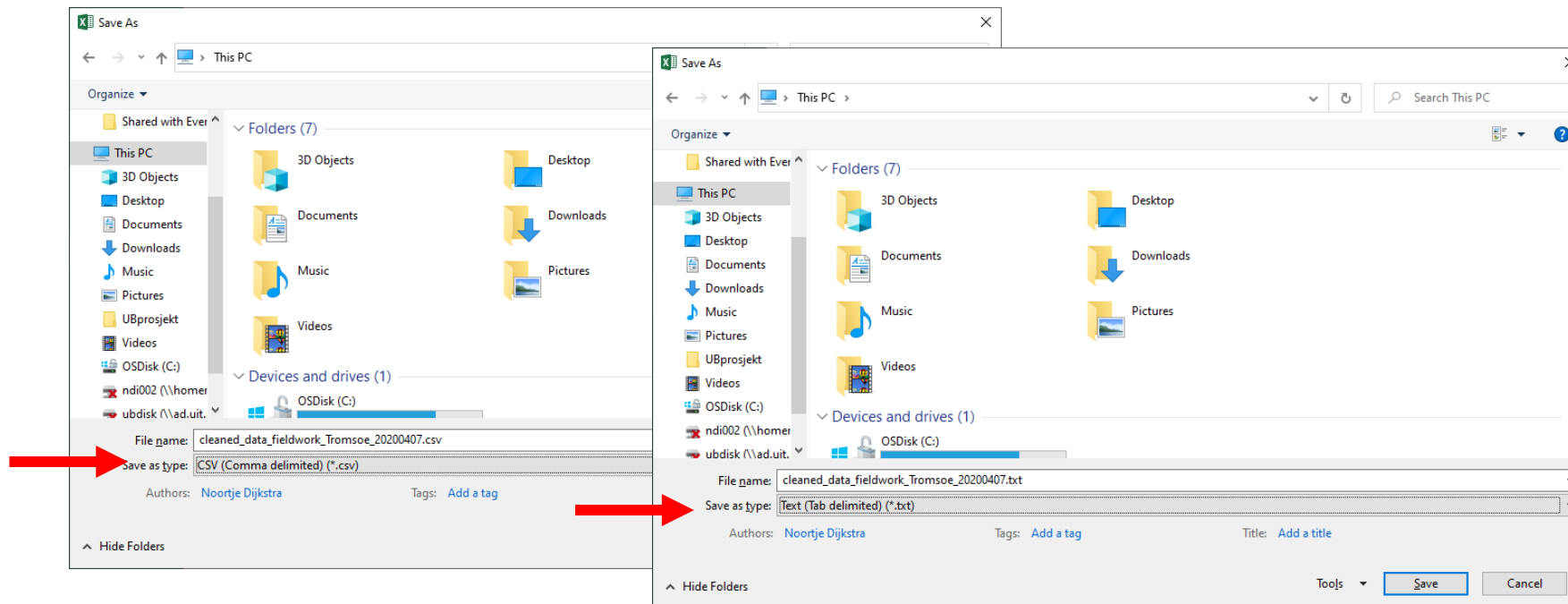
Clear All
Clear Formats
Clear Contents
Clear Comments and Notes
Clear Hyperlinks
Remove Hyperlinks

YEAR	Make	Model	Size	CITY (kWh/100 km)	HWY (kWh/100 km)
2016	NISSAN	LEAF	MID-SIZE	80 16.5	20.8
2016	NISSAN	LEAF	MID-SIZE	80 17	20.7
2015	NISSAN	LEAF	MID-SIZE	80 16.5	20.8
2014	NISSAN	LEAF	MID-SIZE	80 16.5	20.8
2013	NISSAN	LEAF	MID-SIZE	80 19.3	23
2012	NISSAN	LEAF	MID-SIZE	80 19.3	23
2016	MITSUBISHI	i-MIEV	SUBCOMPACT	49 16.9	21.4
2015	MITSUBISHI	i-MIEV	SUBCOMPACT	49 16.9	21.4
2014	MITSUBISHI	i-MIEV	SUBCOMPACT	49 16.9	21.4
2013	MITSUBISHI	i-MIEV	SUBCOMPACT	49 16.9	21.4
2012	MITSUBISHI	i-MIEV	SUBCOMPACT	49 16.9	21.4
2016	TESLA	MODEL X P90D	SUV - STANDARD	568 23.6	23.3
2016	TESLA	MODEL S P85D/P90D	FULL-SIZE	568 23.4	21.5
2016	TESLA	MODEL S P90D (Refresh)	FULL-SIZE	568 22.9	21
2015	TESLA	MODEL S P85D/P90D	FULL-SIZE	515 23.4	21.5
2016	TESLA	MODEL S 70D	FULL-SIZE	386 20.8	20.6
2016	TESLA	MODEL X 90D	SUV - STANDARD	386 23.2	22.2
2016	TESLA	MODEL S 85D/90D	FULL-SIZE	386 22	19.8
2016	TESLA	MODEL S 90D (Refresh)	FULL-SIZE	386 20.8	19.7
2014	TESLA	MODEL S PERFORMANCE	FULL-SIZE	310 23.9	23.2
2013	TESLA	MODEL S PERFORMANCE	FULL-SIZE	310 23.9	23.2
2016	TESLA	MODEL S (60 kWh battery)	FULL-SIZE	283 22.2	21.7
2016	TESLA	MODEL S (70 kWh battery)	FULL-SIZE	283 23.8	23.2
2016	TESLA	MODEL S (85/90 kWh battery)	FULL-SIZE	283 23.8	23.2
2015	TESLA	MODEL S (60 kWh battery)	FULL-SIZE	283 22.2	21.7

	A	B	C	D	E	F	G
1	YEAR	Make	Model3	Size	(kW)	CITY (kWh/100 km)	HWY (kWh/100 km)
2	2016	NISSAN	LEAF	MID-SIZE	80	16.5	20.8
3	2016	NISSAN	LEAF	MID-SIZE	80	17	20.7
4	2015	NISSAN	LEAF	MID-SIZE	80	16.5	20.8
5	2014	NISSAN	LEAF	MID-SIZE	80	16.5	20.8
6	2013	NISSAN	LEAF	MID-SIZE	80	19.3	23
7	2012	NISSAN	LEAF	MID-SIZE	80	19.3	23
8	2016	MITSUBISHI	i-MIEV	SUBCOMPACT	49	16.9	21.4
9	2015	MITSUBISHI	i-MIEV	SUBCOMPACT	49	16.9	21.4
10	2014	MITSUBISHI	i-MIEV	SUBCOMPACT	49	16.9	21.4
11	2013	MITSUBISHI	i-MIEV	SUBCOMPACT	49	16.9	21.4
12	2012	MITSUBISHI	i-MIEV	SUBCOMPACT	49	16.9	21.4
13	2016	TESLA	MODEL X P90D	SUV - STANDARD	568	23.6	23.3
14	2016	TESLA	MODEL S P85D/P90D	FULL-SIZE	568	23.4	21.5
15	2016	TESLA	MODEL S P90D (Refresh)	FULL-SIZE	568	22.9	21
16	2015	TESLA	MODEL S P85D/P90D	FULL-SIZE	515	23.4	21.5
17	2016	TESLA	MODEL S 70D	FULL-SIZE	386	20.8	20.6
18	2016	TESLA	MODEL X 90D	SUV - STANDARD	386	23.2	22.2
19	2016	TESLA	MODEL S 85D/90D	FULL-SIZE	386	22	19.8
20	2016	TESLA	MODEL S 90D (Refresh)	FULL-SIZE	386	20.8	19.7
21	2014	TESLA	MODEL S PERFORMANCE	FULL-SIZE	310	23.9	23.2
22	2013	TESLA	MODEL S PERFORMANCE	FULL-SIZE	310	23.9	23.2
23	2016	TESLA	MODEL S (60 kWh battery)	FULL-SIZE	283	22.2	21.7
24	2016	TESLA	MODEL S (70 kWh battery)	FULL-SIZE	283	23.8	23.2
25	2016	TESLA	MODEL S (85/90 kWh battery)	FULL-SIZE	283	23.8	23.2
26	2015	TESLA	MODEL S (60 kWh battery)	FULL-SIZE	283	22.2	21.7

Export clean data to a text-based format

- Interoperable with most data analysis software programs (I in FAIR data principles)
 - CSV files (.csv) – Consider what separator is used in the cell values (comma, semicolon)
 - TAB delimited (.txt)
 - Export each tab separate.



Source: [Data carpentry](#)



Establish routines for data handling

- Establish a workflow suitable for your studies – and use it consistently!
- Apply descriptive file names and version control to keep track of the workflow.
- Document all changes in a README-file.

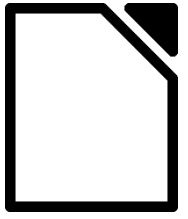
Select the line representing the data
Right click the line
Select – Format Data Series
Select – Line Style
Check – the box for Smoothed line

Select the line representing the data
Click again on data point to be edited
Right click data point to be edited
Select – Format Data Point
Select – Marker Options
Select – Built-in
For Type, Select – the circle
For Size, Select – 8

Select – Marker Fill
Select – Solid Fill
Select – Color (the paint bucket)
Select – White
Select – Marker Line Color
Select – Solid Line
Select – Color (the paint bucket)
Select – the same color you used for the main line
Select – Marker Line Style
For Width, increase to 2 pt

Spreadsheets

Excel, Google Sheets, LibreOffice

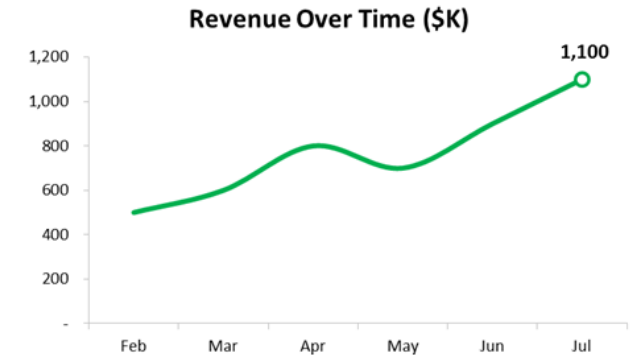
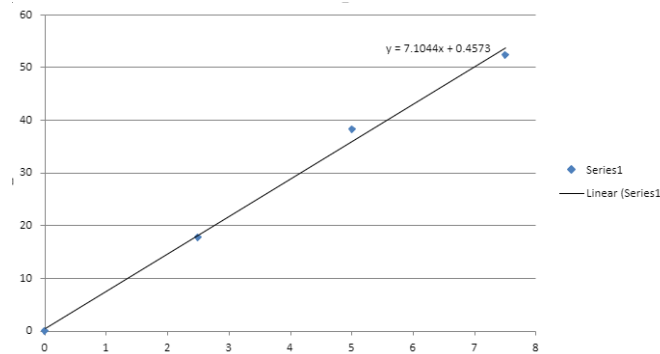


Pros:

- Available, easy to use, quick overview of the data

Cons:

- Black box; lack of control
- Difficult to track changes and reproduce the steps



Solution: Use Excel [Macros](#) to record the workflow

Tools for data cleaning

Programming languages Python or R

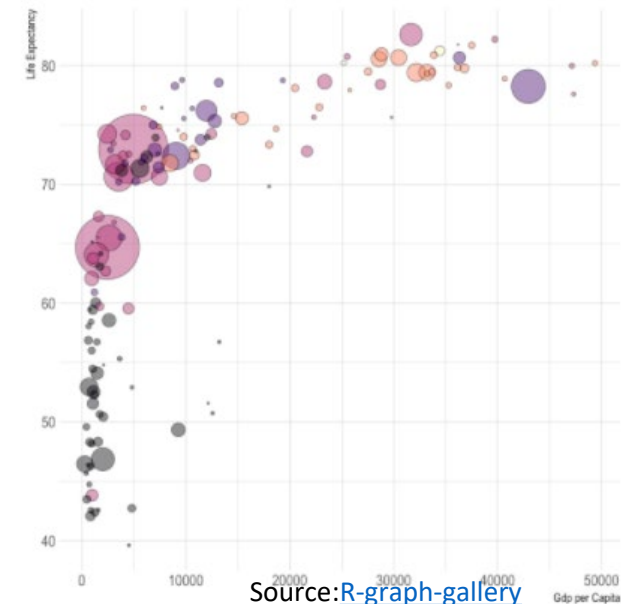


Pros:

- Free, open-source, works across platforms (FAIR-principles: Accessible and Interoperable)
- Full control of data processing
- Easy to track changes
- Reproducible!
- Help online (e.g. [Stack Overflow](#), [RStudio community](#), [Python community](#))
- High quality visualizations

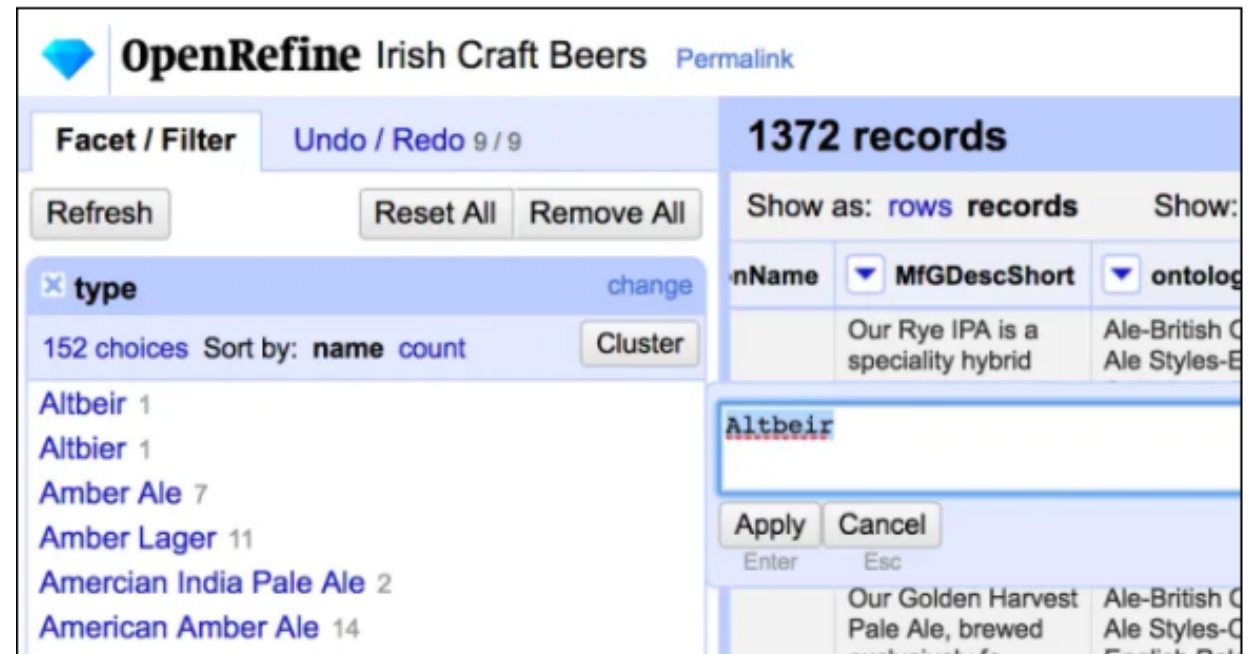
Cons:

- Need to learn the language



Datacleaning with OpenRefine

- Overview of large datasets
- Easy to identify and fix errors and irregularities.
- Easy to combine data from different sources.
- OpenRefine does not change the original file.
- All actions are reversible
- All actions are tracked, and the documentation can be published alongside the data.
- The workflow can be saved and applied to new datasets.



Recommendation for self-study

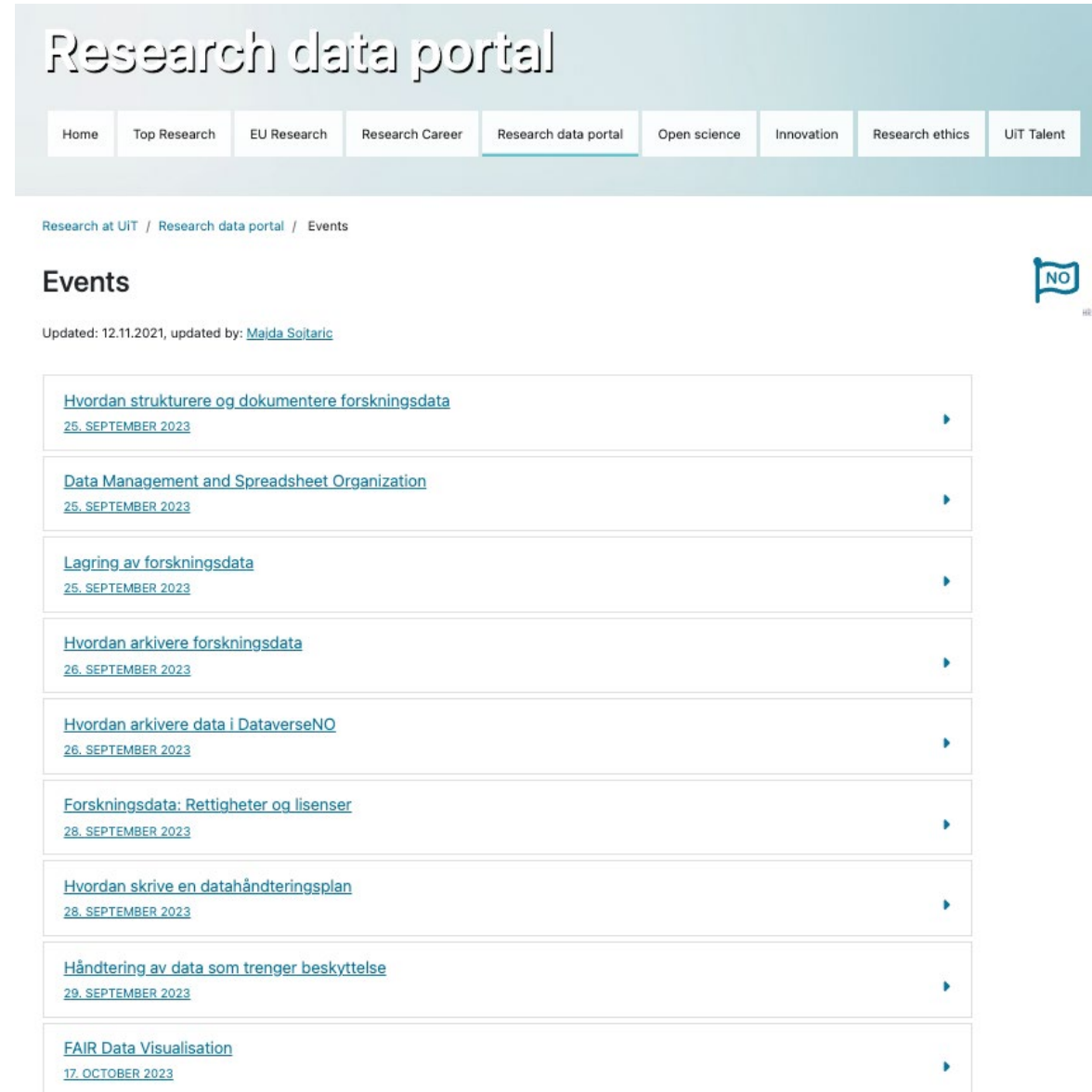
How to clean your data and apply a quality control (Data Carpentry lessons)

- [Data Organization in Spreadsheets for Ecologists](#)
- [Data Cleaning with OpenRefine for Ecologists](#)
- [Data Organization in Spreadsheets for Social Scientists](#)
- [OpenRefine for Social Science Data](#)

Upcoming webinars on Research Data Management

Courses at UiT are open to anyone. For more info and registration, click [here](#)

<https://en.uit.no/research/research-dataportal>

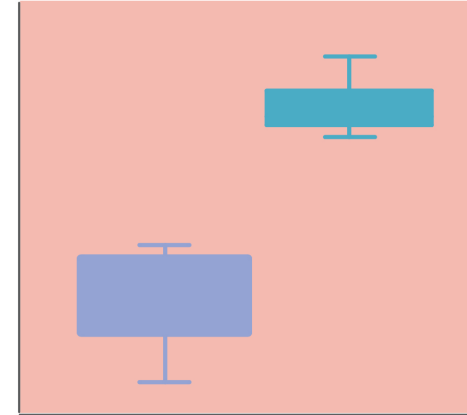
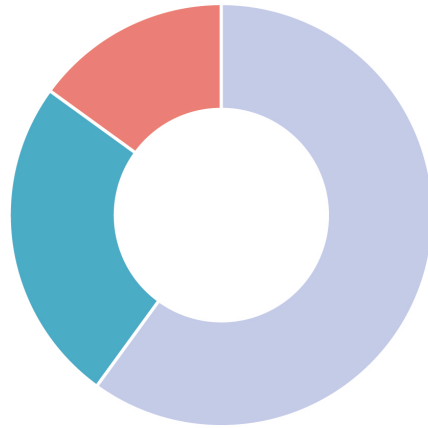


The screenshot shows the 'Research data portal' website. At the top, there is a navigation menu with links for Home, Top Research, EU Research, Research Career, Research data portal (highlighted), Open science, Innovation, Research ethics, and UiT Talent. Below the menu, the breadcrumb trail reads 'Research at UiT / Research data portal / Events'. The main heading is 'Events', with a small 'NO' flag icon and 'H5 EX' text to the right. Below the heading, it says 'Updated: 12.11.2021, updated by: [Majda Soitaric](#)'. A list of ten webinars is displayed, each with a title, date, and a right-pointing arrow:

- [Hvordan strukturere og dokumentere forskningsdata](#)
25. SEPTEMBER 2023
- [Data Management and Spreadsheet Organization](#)
25. SEPTEMBER 2023
- [Lagring av forskningsdata](#)
25. SEPTEMBER 2023
- [Hvordan arkivere forskningsdata](#)
26. SEPTEMBER 2023
- [Hvordan arkivere data i DataverseNO](#)
26. SEPTEMBER 2023
- [Forskningsdata: Rettigheter og lisenser](#)
28. SEPTEMBER 2023
- [Hvordan skrive en datahåndteringsplan](#)
28. SEPTEMBER 2023
- [Håndtering av data som trenger beskyttelse](#)
29. SEPTEMBER 2023
- [FAIR Data Visualisation](#)
17. OCTOBER 2023



Upcoming webinars on Research Data Management



Courses at UiT are open to anyone. For more info and registration, click [here](#)

FAIR Data Visualisation @ UiT

October 17-23, 2023

Realise the potential of your research data.

This 4-day workshop includes lectures on FAIR data; principles of data visualisation; charts & attributes; oral presentation skills; & an introduction to Python and Jupyter notebooks. Participants will construct and present visualisations of their own research data.



Register on Tavla or contact:

Katie Smart *The University Library (UB)*
kathleen.a.smart@uit.no

Radovan Bast *Research Software Engineering (IT)*
radovan.bast@uit.no



UiT Data Stewards Network (DSN)

Collaboratively building a culture of best practices for research data management

What is a Data Steward ?

Custodian of data within research or lab groups, or at the institute-level.

Ensures data is collected and managed according to best practises.

Can include managing physical samples, archives, collections.



UiT Data Stewards Network (DSN)

Collaboratively building a culture of best practices for research data management

- Create an active community for networking and professional development
- Collaboration and knowledge sharing across disciplines
- Support researchers and research administrators
- Promote the FAIR principles and research data reuse
- Ensure ethical and responsible management of qualitative and sensitive data
- Engage faculties and research administration for consolidated support
- Provide resources and incentives promoting diligence in research data management

Activities: two meetings per semester with topics to be announced on Tavla

Communication platform: Teams channel [UiT Data Stewards Network](#)

Contact: researchdata@uit.no



Information and help



[Research Data Management at UiT](#)



Email: researchdata@hjelp.uit.no

References

Teal et al., 2019, datacarpentry/spreadsheet-ecology-lesson: Data Carpentry: Data Organization in Spreadsheets for Ecologists, June 2019: Zenodo, doi:10.5281/zenodo.3269869.

White et al., 2013. Nine simple ways to make it easier to (re)use your data, Ideas in Ecology and Evolution 6(2): 1-10 Special Issue-Data Sharing in Ecology and Evolution
<https://ojs.library.queensu.ca/index.php/IEE/article/view/4608>

Hadley Wickham, *Tidy Data*, Vol. 59, Issue 10, Sep 2014, Journal of Statistical Software. <http://www.jstatsoft.org/v59/i10>.