ACENET

accelerate discovery

CERC Health Equity & Community Wellbeing Training

January 29-30, 2025

Led by: Meghan Landry, ACENET

A regional partner of the
Digital Research Alliance of Canada

The following lesson has been adapted from the Library Carpentry’s Tidy Data lesson. These materials are free to share and adapt as needed.

Lesson objectives

What we will cover today:

- Good **data entry** & **organization** practices, such as formatting data tables in spreadsheets
- How to avoid common formatting mistakes
- Dates as data (beware!)
- Basic quality control and data manipulation in spreadsheets
- Exporting data from spreadsheets

What we will *not* cover today:

- How to do statistics in a spreadsheet
- How to do plotting in a spreadsheet
- How to write code in spreadsheet programs
- **Data analysis**

Using spreadsheets programs for data organization

Good data organization is the foundation of much of our day-to-day work as academics

What we usually use spreadsheets for:

- **Data entry**
- **Organizing data**
- **Subsetting and sorting data**
- **Statistics**
- **Plotting**, and more

Much of your time when you're producing a report will be spent in this 'data wrangling' stage. It's not the most fun, but it's necessary.

A quick note about spreadsheet programs

There are a number of programs to use on a desktop or web browser:

- LibreOffice Calc (free)
- Microsoft Excel
- Apple Numbers
- Google Sheets
- Gnumeric
- Apache OpenOffice Calc

Things to keep in mind:

- Commands may differ between programs, between operating systems (i.e. Excel looks different on Mac vs. Windows), and between versions
- My interface may look different than yours and **that’s okay**!

Problems with spreadsheets

- Spreadsheets are good for **data entry**, but the reality is that we use it for much more than that: creating data tables, generating summary statistics, and making figures
- Spreadsheet programs **are not suitable for many tasks and jobs**, and we need to consider when to use other programs for our specific tasks or goals
- Spreadsheet programs can *sometimes* be suitable for **data cleaning**, prior to importing data into a statistical analysis program.
    - We will discuss data cleaning software, such as OpenRefine, at the end of the presentation

Questions to ask yourself about your data or project

When to use Excel? When you:

- Want a **flat** or **non relational view** of your data
- When your data is mostly **numeric**
- Frequently run **calculations & statistical comparisons** on your data
- Want to perform sophisticated **what-if analysis operations** on your data
- Want to keep track of items in a **simple list**, either for personal use or for limited collaboration purposes

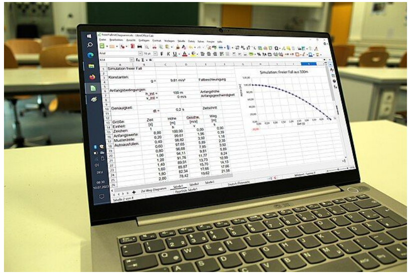
When to use a database (Access, SQL)? When you:

- Anticipate **many people working in the database**
- Anticipate the **need to add more tables to a dataset** that originated as flat table
- Want to **run complex queries**
- Want to **produce a variety of reports**

Adapted from "Using Access or Excel to manage your data"

Formatting data tables in spreadsheets

Well-formatted tables and data organization is the key



- The most common mistake we make is treating a spreadsheet like **it is a notebook** by relying on:
  - Context
  - Notes in the margin
  - Spatial or visual layout of data & fields to convey information
  - And making it look ‘pretty’

Keeping track of your analyses

1. **Create a new file or tab** with your cleaned or analyzed data. Do not modify that original dataset, or you will never know where you started!

2. **Keep track of the steps you took** in your clean up or analysis. You can do this by:
   a. Creating a new text file
   b. Create a new tab in your spreadsheet with your notes
      i. Save your spreadsheet with a **file format compatible with multiple tabs**, if you do this

Structuring data in spreadsheets

1. Put all your **variables in columns** - the thing you're measuring, like 'length' or 'attendance'

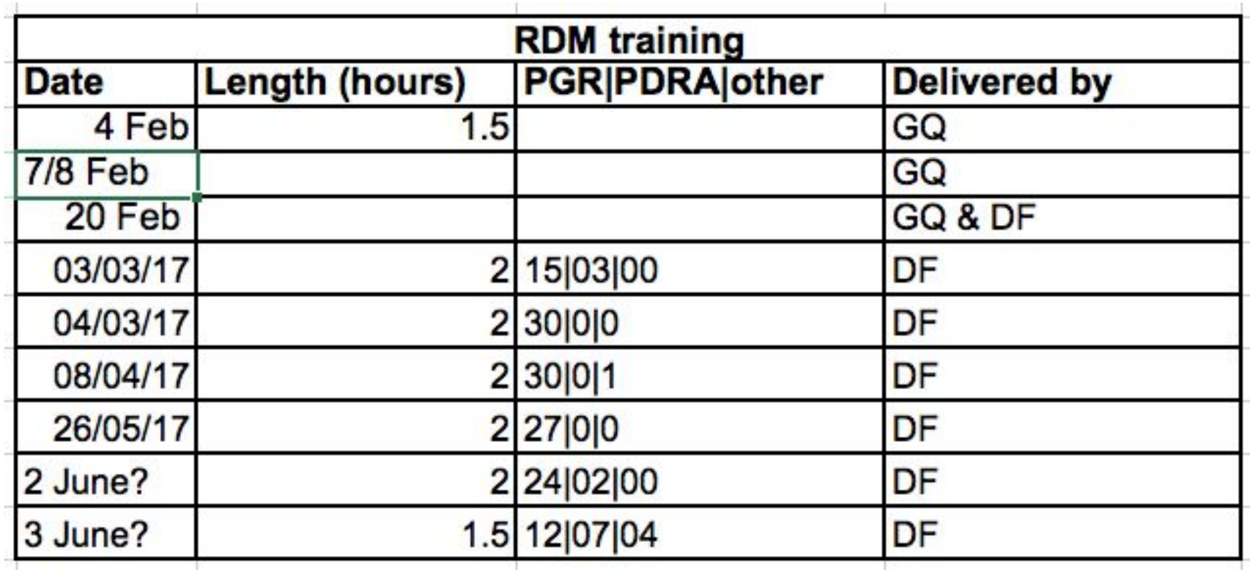
2. Put each **observation in its own row**

3. **Don't combine multiple pieces of information** in one cell

4. Leave the raw data, **raw**

5. Export the cleaned data to a **text-based format like CSV**

If you were to keep track of data like this..



<table>
<tr><th colspan="4">RDM training</th></tr>
<tr><th>Date</th><th>Length (hours)</th><th>PGR|PDRA|other</th><th>Delivered by</th></tr>
<tr><td>4 Feb</td><td>1.5</td><td></td><td>GQ</td></tr>
<tr><td>7/8 Feb</td><td></td><td></td><td>GQ</td></tr>
<tr><td>20 Feb</td><td></td><td></td><td>GQ &amp; DF</td></tr>
<tr><td>03/03/17</td><td>2</td><td>15|03|00</td><td>DF</td></tr>
<tr><td>04/03/17</td><td>2</td><td>30|0|0</td><td>DF</td></tr>
<tr><td>08/04/17</td><td>2</td><td>30|0|1</td><td>DF</td></tr>
<tr><td>26/05/17</td><td>2</td><td>27|0|0</td><td>DF</td></tr>
<tr><td>2 June?</td><td>2</td><td>24|02|00</td><td>DF</td></tr>
<tr><td>3 June?</td><td>1.5</td><td>12|07|04</td><td>DF</td></tr>
</table>

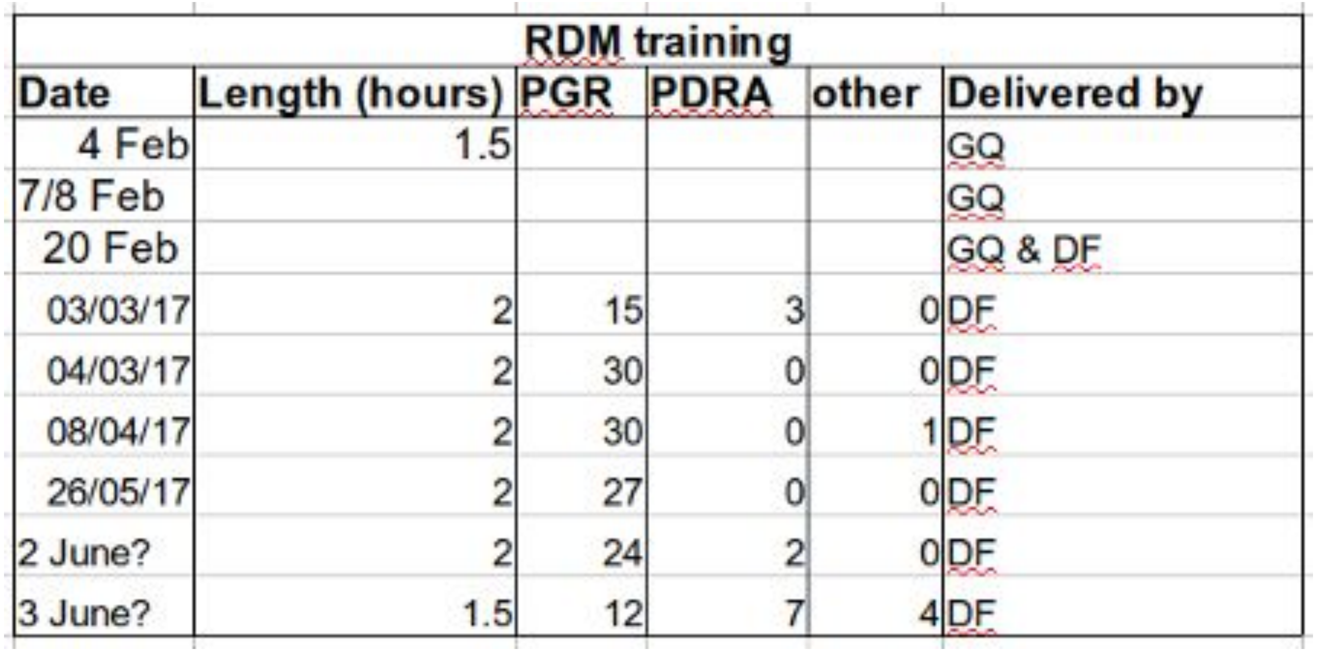
The problem?

The number of attendees of different types (post-grad researcher (PGR), post-doc research associate (PDRA), & other) are in the same field.

The solution?

Put attendee categories in different columns.

Columns for variables, rows for observations



RDM training					
Date	Length (hours)	PGR	PDRA	other	Delivered by
4 Feb	1.5				GQ
7/8 Feb					GQ
20 Feb					GQ & DF
03/03/17	2	15	3	0	DF
04/03/17	2	30	0	0	DF
08/04/17	2	30	0	1	DF
26/05/17	2	27	0	0	DF
2 June?	2	24	2	0	DF
3 June?	1.5	12	7	4	DF

Columns = **variables**

Rows = **observations**

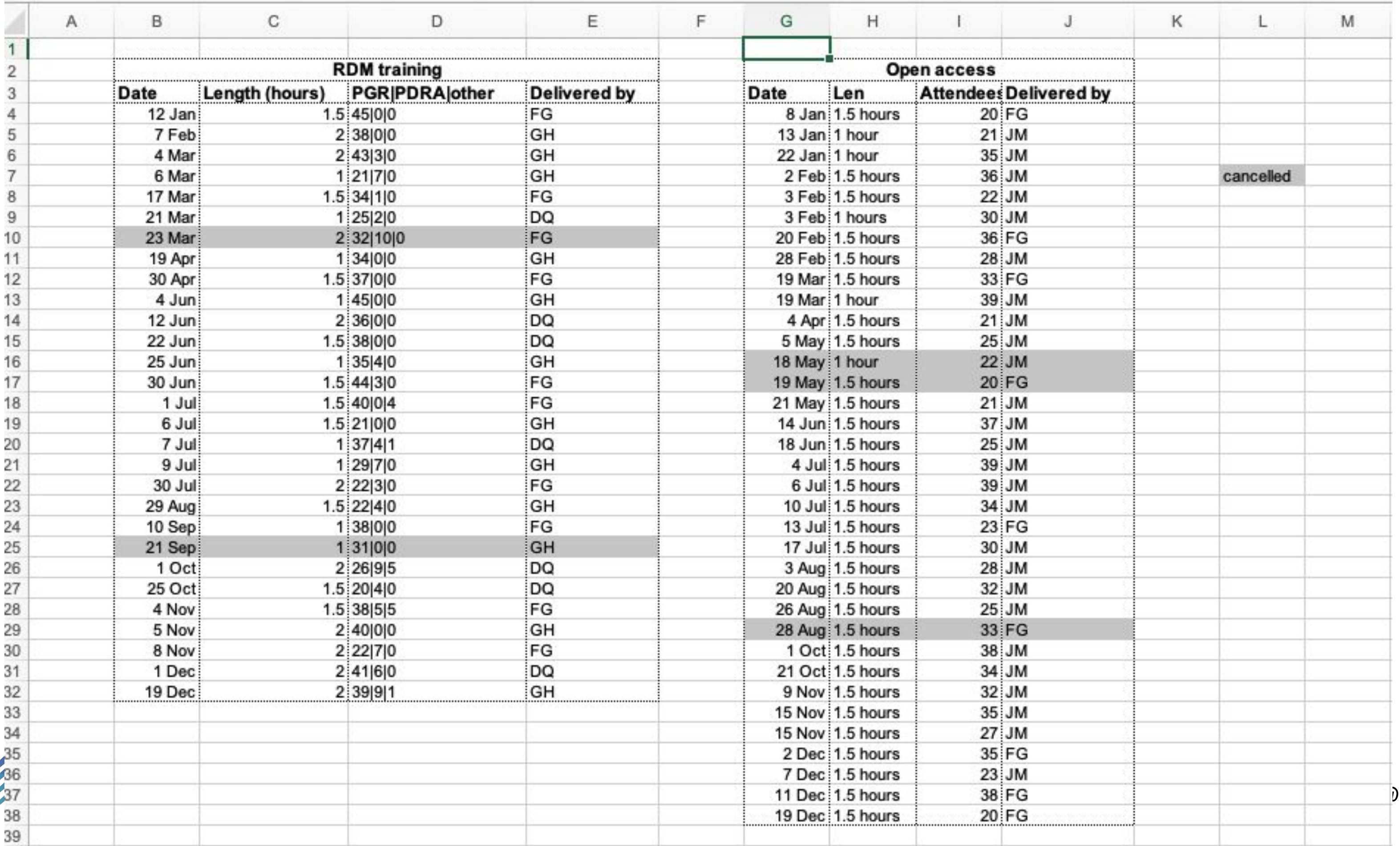
Cells = **data (values)**

Formatting problems

Common spreadsheet errors

1. Multiple tables
2. Multiple tabs
3. Not filling in zeros
4. Using bad null values
5. Using formatting to convey information
6. Using formatting to make the data sheet look pretty
7. Placing comments or units in cells
8. More than one piece of information in a cell
9. Field name problems
10. Special characters in data
11. Inclusion of metadata in data table

Multiple tables



RDM training

Date	Length (hours)	PGR|PDRA|other	Delivered by
12 Jan	1.5	45|0|0	FG
7 Feb	2	38|0|0	GH
4 Mar	2	43|3|0	GH
6 Mar	1	21|7|0	GH
17 Mar	1.5	34|1|0	FG
21 Mar	1	25|2|0	DQ
23 Mar	2	32|10|0	FG
19 Apr	1	34|0|0	GH
30 Apr	1.5	37|0|0	FG
4 Jun	1	45|0|0	GH
12 Jun	2	36|0|0	DQ
22 Jun	1.5	38|0|0	DQ
25 Jun	1	35|4|0	GH
30 Jun	1.5	44|3|0	FG
1 Jul	1.5	40|0|4	FG
6 Jul	1.5	21|0|0	GH
7 Jul	1	37|4|1	DQ
9 Jul	1	29|7|0	GH
30 Jul	2	22|3|0	FG
29 Aug	1.5	22|4|0	GH
10 Sep	1	38|0|0	FG
21 Sep	1	31|0|0	GH
1 Oct	2	26|9|5	DQ
25 Oct	1.5	20|4|0	DQ
4 Nov	1.5	38|5|5	FG
5 Nov	2	40|0|0	GH
8 Nov	2	22|7|0	FG
1 Dec	2	41|6|0	DQ
19 Dec	2	39|9|1	GH

Open access

Date	Len	Attendees	Delivered by
8 Jan	1.5 hours	20	FG
13 Jan	1 hour	21	JM
22 Jan	1 hour	35	JM
2 Feb	1.5 hours	36	JM
3 Feb	1.5 hours	22	JM
3 Feb	1 hours	30	JM
20 Feb	1.5 hours	36	FG
28 Feb	1.5 hours	28	JM
19 Mar	1.5 hours	33	FG
19 Mar	1 hour	39	JM
4 Apr	1.5 hours	21	JM
5 May	1.5 hours	25	JM
18 May	1 hour	22	JM
19 May	1.5 hours	20	FG
21 May	1.5 hours	21	JM
14 Jun	1.5 hours	37	JM
18 Jun	1.5 hours	25	JM
4 Jul	1.5 hours	39	JM
6 Jul	1.5 hours	39	JM
10 Jul	1.5 hours	34	JM
13 Jul	1.5 hours	23	FG
17 Jul	1.5 hours	30	JM
3 Aug	1.5 hours	28	JM
20 Aug	1.5 hours	32	JM
26 Aug	1.5 hours	25	JM
28 Aug	1.5 hours	33	FG
1 Oct	1.5 hours	38	JM
21 Oct	1.5 hours	34	JM
9 Nov	1.5 hours	32	JM
15 Nov	1.5 hours	35	JM
15 Nov	1.5 hours	27	JM
2 Dec	1.5 hours	35	FG
7 Dec	1.5 hours	23	JM
11 Dec	1.5 hours	38	FG
19 Dec	1.5 hours	20	FG

cancelled

Multiple tabs

When you create extra tabs, you fail to allow the computer to see connections in the data. Say you make a separate tab for each year, this is **bad practice for 2 reasons:**

- You are more likely to add inconsistencies to your data if each time you take a measurement, you record it in a new tab
- You will add an extra step for yourself before you analyze the data because you will have to combine these data into a single data table.

Ask yourself: ***"Self, could I avoid adding this tab by adding another column to my original spreadsheet?"***

This can get long, so you can **freeze column headers**:

Documentation on how to freeze column headers in Microsoft Excel
Documentation on how to freeze column headers in LibreOffice Calc
Documentation on how to freeze column headers in Google Sheets

Not filling in zeroes

- There’s a difference between **a zero** and **a blank cell** in a spreadsheet.
    - To the computer, a zero is actually data - you measured or counted it.
    - A blank cell means that it wasn’t measured and the computer will interpret it as a null value.

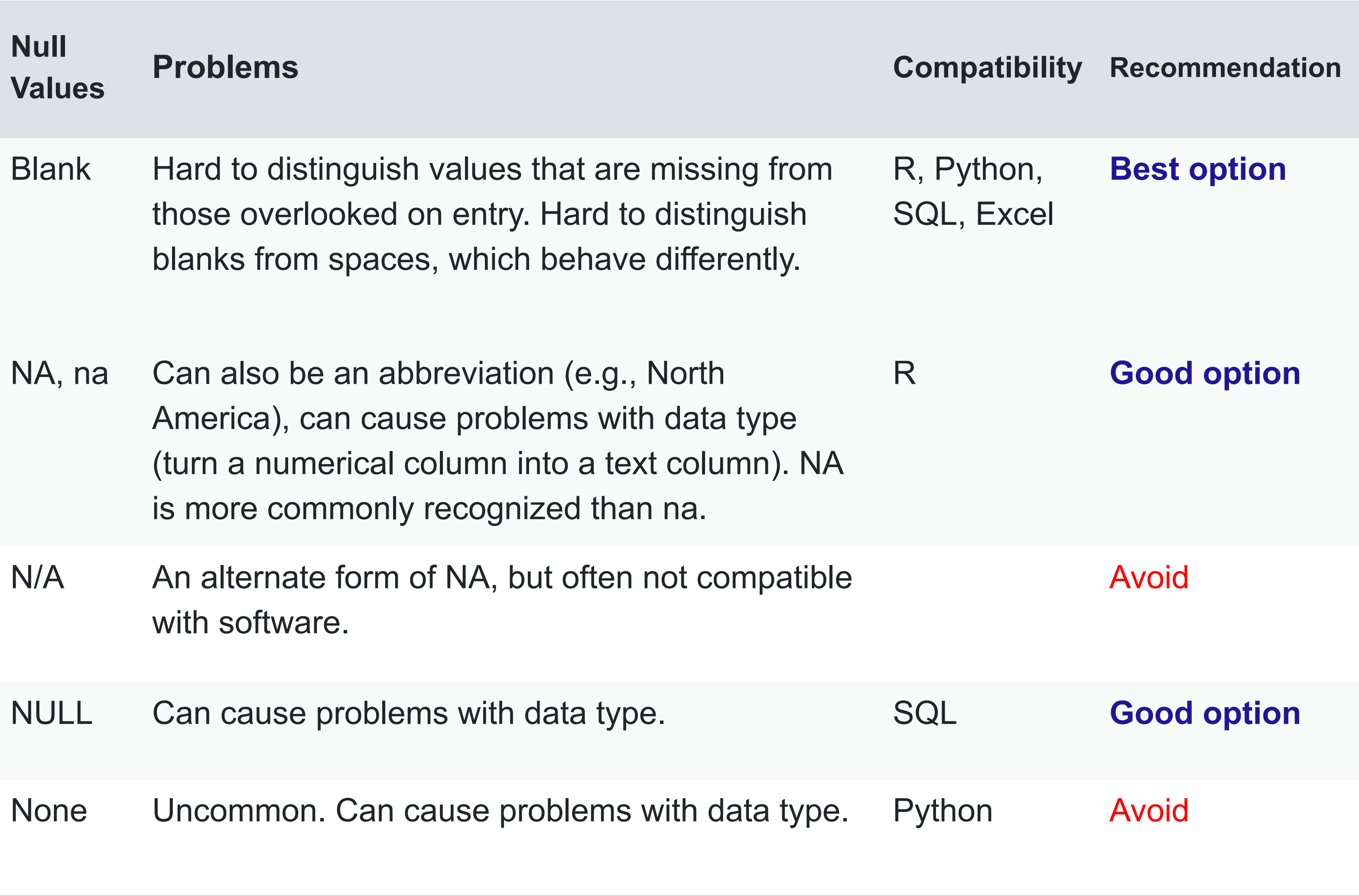
- Spreadsheets or statistical programs will likely misinterpret blank cells that are meant to be zero - this is the equivalent to leaving out data. **Zero observations are real data!**

Using bad null values

Example: using `-999`, other numerical values, zero, or text to represent missing values. Whatever the reason, it's a problem if unknown or missing data is recorded as `-999`, `999`, or `0`. Many statistical programs will not recognize that these are intended to represent missing (null) values. How these values are interpreted will depend on the software you use to analyze your data.

Solution: A solution will depend on the final application of your data and how you intend to analyse it, but it is essential to use a clearly defined and **consistent** null indicator. Blank cells are the best choices for most applications; when working in R, NA may be an acceptable null value choice.

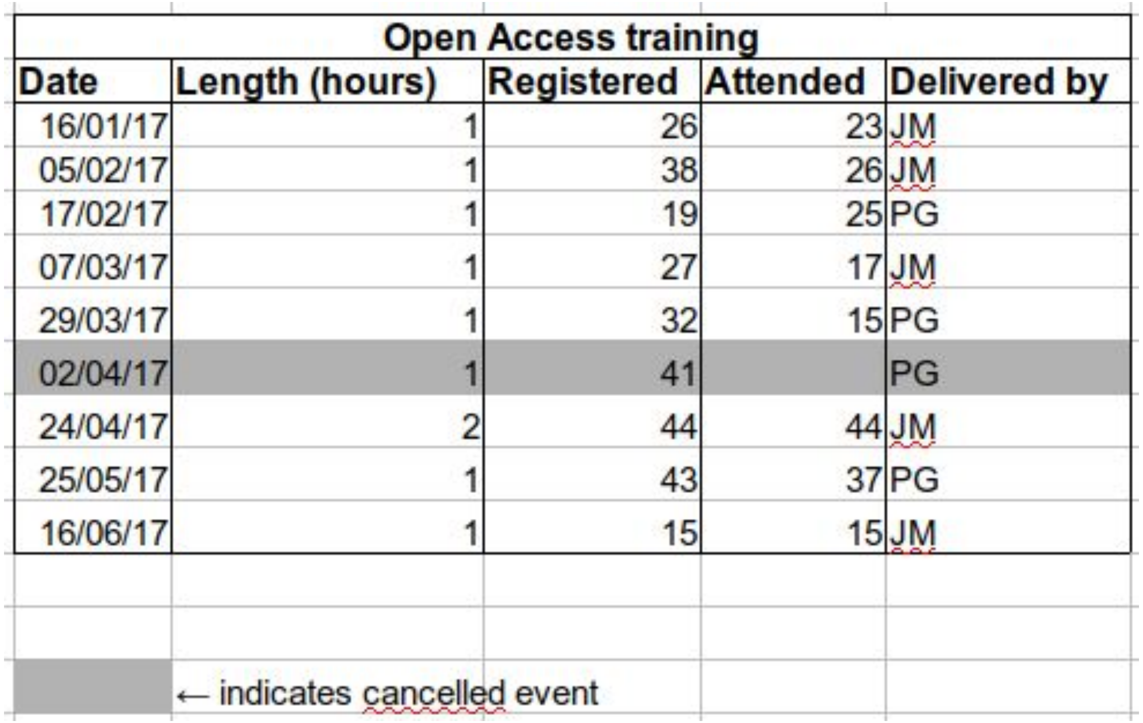
If null values convey different reasons why the data is missing, consider creating a new column like `data_missing` to capture the reasons for missing data!



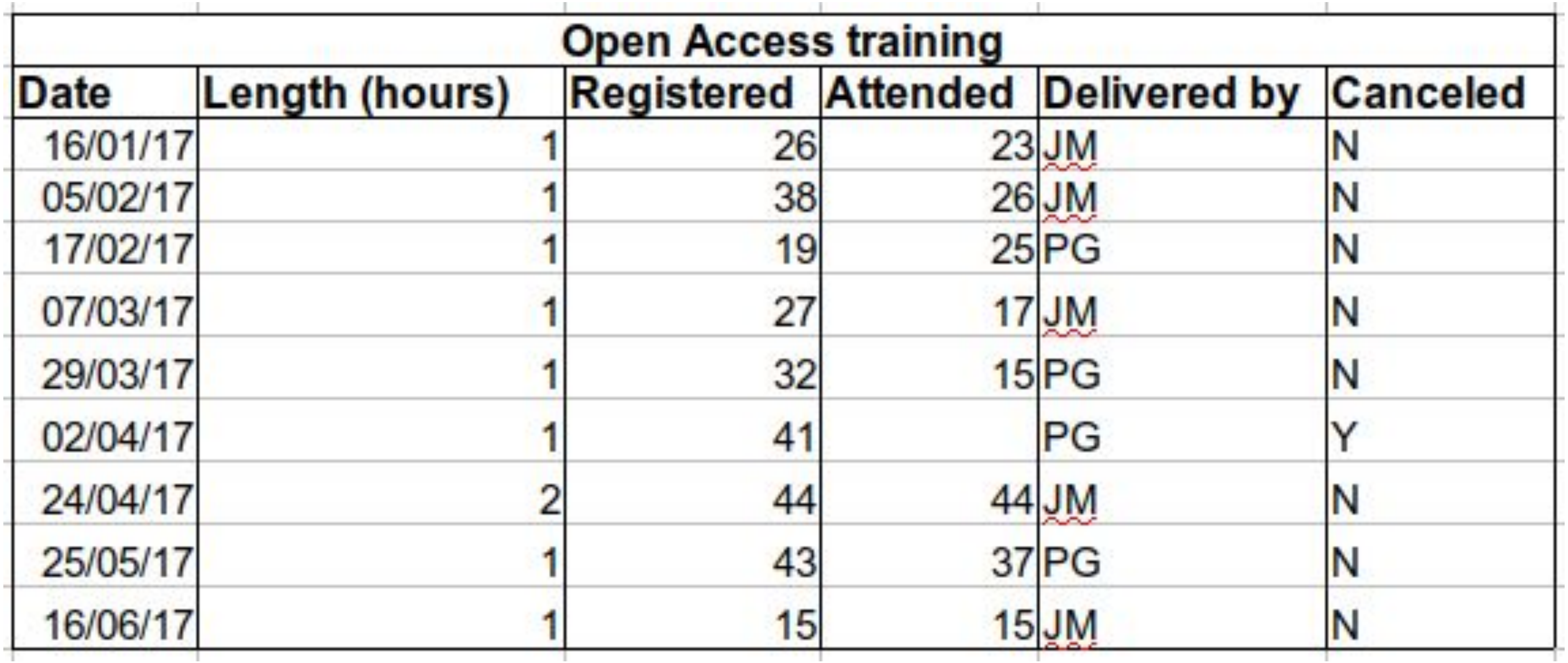
Null Values	Problems	Compatibility	Recommendation
Blank	Hard to distinguish values that are missing from those overlooked on entry. Hard to distinguish blanks from spaces, which behave differently.	R, Python, SQL, Excel	**Best option**
NA, na	Can also be an abbreviation (e.g., North America), can cause 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

Using formatting to convey information

Example: highlighting cells, rows or columns that should be excluded from an analysis, leaving blank rows to indicate separations in data.



Open Access training				
Date	Length (hours)	Registered	Attended	Delivered by
16/01/17	1	26	23	JM
05/02/17	1	38	26	JM
17/02/17	1	19	25	PG
07/03/17	1	27	17	JM
29/03/17	1	32	15	PG
02/04/17	1	41		PG
24/04/17	2	44	44	JM
25/05/17	1	43	37	PG
16/06/17	1	15	15	JM
▨	← indicates cancelled event			



Open Access training

Date	Length (hours)	Registered	Attended	Delivered by	Canceled
16/01/17	1	26	23	JM	N
05/02/17	1	38	26	JM	N
17/02/17	1	19	25	PG	N
07/03/17	1	27	17	JM	N
29/03/17	1	32	15	PG	N
02/04/17	1	41		PG	Y
24/04/17	2	44	44	JM	N
25/05/17	1	43	37	PG	N
16/06/17	1	15	15	JM	N

Solution: create a new field to encode which data should be excluded.

Using formatting to make the data sheet look pretty

Example: merging cells.

Solution: If you're not careful, formatting a worksheet to be more aesthetically pleasing can compromise your computer's ability to see associations in the data. Merged cells are an absolute formatting NO-NO if you want to make your data readable by statistics software. **Consider restructuring your data in such a way that you will not need to merge cells to organize your data**.

Placing comments or units in cells

Example: Your data was collected, in part, by a summer student who you later found out was mis-recording the duration of training sessions, some of the time. You want a way to note these data are suspect.

Solution: Most statistical programs can’t see Excel’s comments, and would be confused by comments placed within your data cells. As described above for formatting, **create another field if you need to add notes to cells**.

- Similarly, don’t include units in cells (such as “hours”,“min”): **ideally, all the units or measurements you place in one column should be of the same standard, but if for some reason they aren’t, insert another column and specify the units.**

More than one piece of information in a cell

Example: One table recorded attendance by the different types of attendees. This table recorded the number of attendees of different types: post-graduate researcher (PGR), postdoctoral research associate (PDRA), and others.

Solution: Never include more than one piece of information in a cell. **Design your datasheet to include a column for each type of attendee, if this information is important to collect, rather than just a total number.**

Field name problems

Choose descriptive field names, but be careful not to include: **spaces, numbers, or special characters of any kind**.

- **Spaces** can be misinterpreted by parsers that use whitespace as delimiters and some programs don’t like field names that are text strings that start with numbers.
- **Underscores (_)** are a good alternative to spaces and consider writing names in `CamelCase` to improve readability.
- Remember that abbreviations that make sense at the moment may not be so obvious in 6 months.
- **Including the units in the field names avoids confusion** and enables others to readily interpret your fields.

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
length	length	l
cell_type	CellType	Cell Type
Observation_01	first_observation	1st Obs

Special characters in data

Example: You treat Excel as a word processor when writing notes, even copying data directly from Word or other applications.

Solution: This is a common strategy. For example, when writing longer text in a cell, people often include line breaks, em-dashes (--), et al in their spreadsheet. When exporting this data into a coding/statistical environment or into a relational database, dangerous things may occur, such as lines being cut in half and encoding errors being thrown.

General best practice is to avoid adding characters such as newlines, tabs, and vertical tabs. In other words, treat a text cell as if it were a simple web form that can only contain text and spaces.

Inclusion of metadata in data table

Example: You add a legend at the top or bottom of your data table explaining column meaning, units, exceptions, etc.

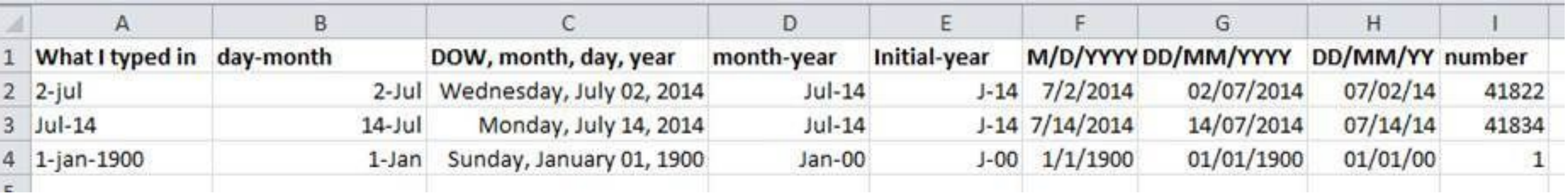
Solution: While recording data about your data (“metadata”) is essential, this information should not be contained in the data file itself. Unlike a table in a paper or a supplemental file, metadata (in the form of legends) should not be included in a data file since this information is not data, and including it can disrupt how computer programs interpret your data file.

- **Rather, metadata should be stored as a separate file in the same directory as your data file, preferably in plain text format with a name that clearly associates it with your data file**.

Dates as data

- Storing dates in one column is not a good practice.
- Spreadsheet programs may display the dates incorrectly (for readability) but how it actually stores the dates may be problematic.
- Date information may be changed when data is converted to different spreadsheet formats, such as between `.xlsx` and `.csv`

Date formats in spreadsheets



	A	B	C	D	E	F	G	H	I
1	What I typed in	day-month	DOW, month, day, year	month-year	Initial-year	M/D/YYYY	DD/MM/YYYY	DD/MM/YY	number
2	2-jul	2-Jul	Wednesday, July 02, 2014	Jul-14	J-14	7/2/2014	02/07/2014	07/02/14	41822
3	Jul-14	14-Jul	Monday, July 14, 2014	Jul-14	J-14	7/14/2014	14/07/2014	07/14/14	41834
4	1-jan-1900	1-Jan	Sunday, January 01, 1900	Jan-00	J-00	1/1/1900	01/01/1900	01/01/00	1

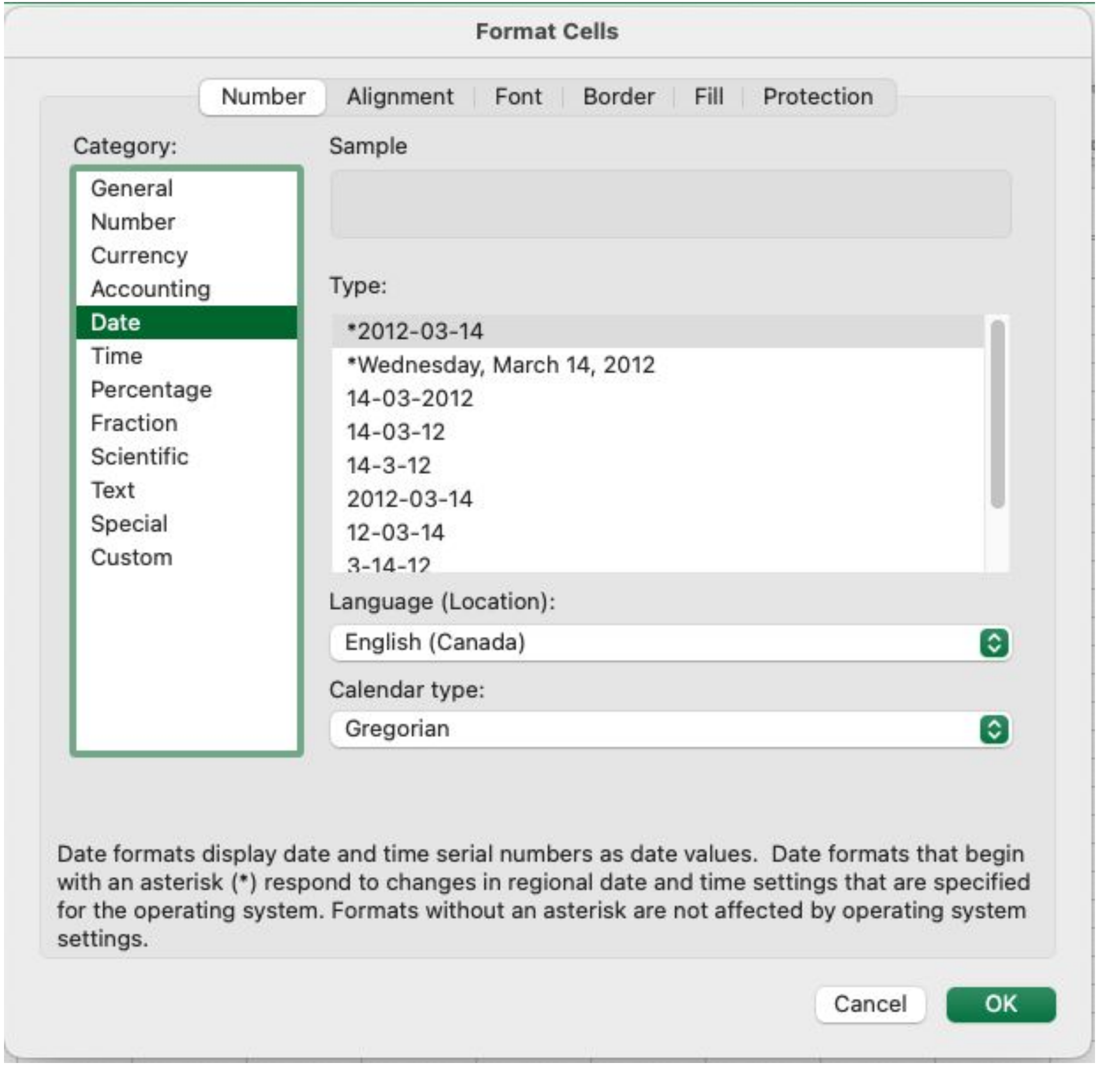
Displaying dates

Ambiguity may creep into your data in numerous ways depending on the format you choose when you enter your data. You may find that Excel will interpret your data in unexpected ways later.

The display format of each cell can be modified.

To change the display in Excel:

- Navigate to the **Format** menu and choose "**Cells...**"
- In the "**Format Cells**" dialog box, you can select a Date format and choose various display outputs.
- In the dialog box, you can also choose to format the cell as a **number** or **text**. It may be useful to format as one or the other because spreadsheet programs understands date information as a number..



Category:

General
Number
Currency
Accounting
Date
Time
Percentage
Fraction
Scientific
Text
Special
Custom

Type:

*2012-03-14
*Wednesday, March 14, 2012
14-03-2012
14-03-12
14-3-12
2012-03-14
12-03-14
3-14-12

Language (Location): English (Canada)

Calendar type: Gregorian

Date formats display date and time serial numbers as date values. Date formats that begin with an asterisk (*) respond to changes in regional date and time settings that are specified for the operating system. Formats without an asterisk are not affected by operating system settings.

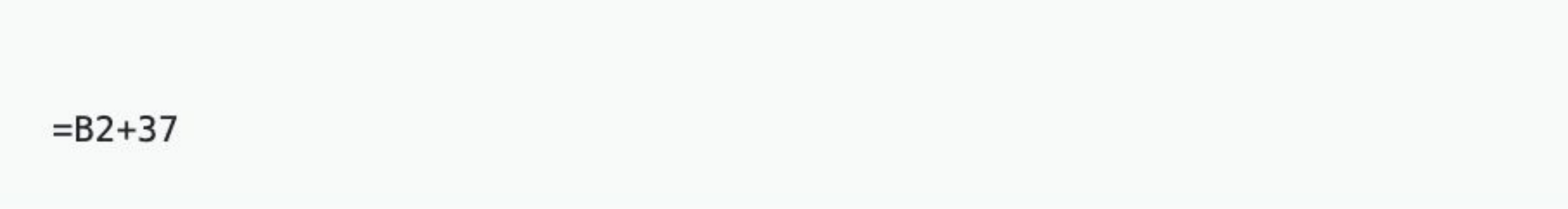
Storing dates

Spreadsheet applications, including Excel, store dates as a **number**.

Developers chose a single day to designate as day zero, and each subsequent day is incremented by a value of one. For example, Excel counts the days from a default of **December 31, 1899**. Thus, **July 2, 2014** is stored as the serial number `41822` because it is 41,822 days after day zero.

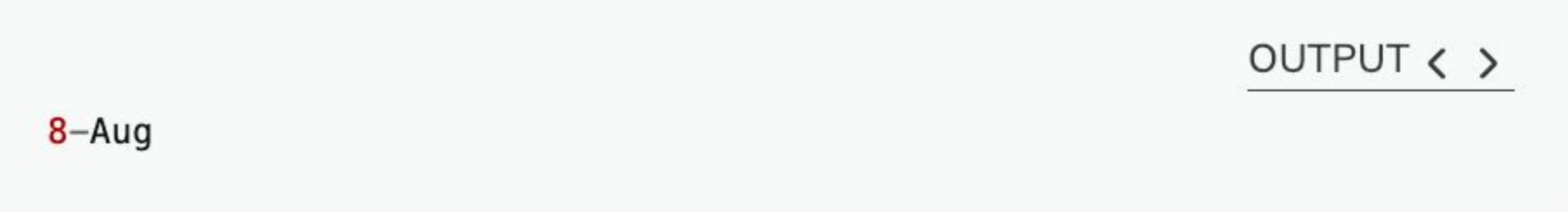
Understanding that these programs use serial numbers to process dates can be useful – you can easily add days, months, or years to a given date.

Say you had a sampling plan where you need to sample every 37 days..



```
=B2+37
```

This would display:



```
8–Aug
```

Working with historical dates

Excel is unable to parse dates from before 1899-12-31, and will thus leave these untouched. **If you're mixing historic data from before and after this date, Excel will translate only the post-1900 dates into its internal format, thus resulting in mixed data.** If you're working with historic data, be extremely careful with your dates!

Excel also entertains a second date system, **the 1904 date system**, as the default in Excel for Macintosh. This system will assign a different serial number than the 1900 date system. Because of this, dates must be checked for accuracy when exporting data from Excel (look for dates that are about 4 years off).

Useful spreadsheet functions for working with date information

If a date is entered in one column, we can use functions to extract information from that column into other columns.

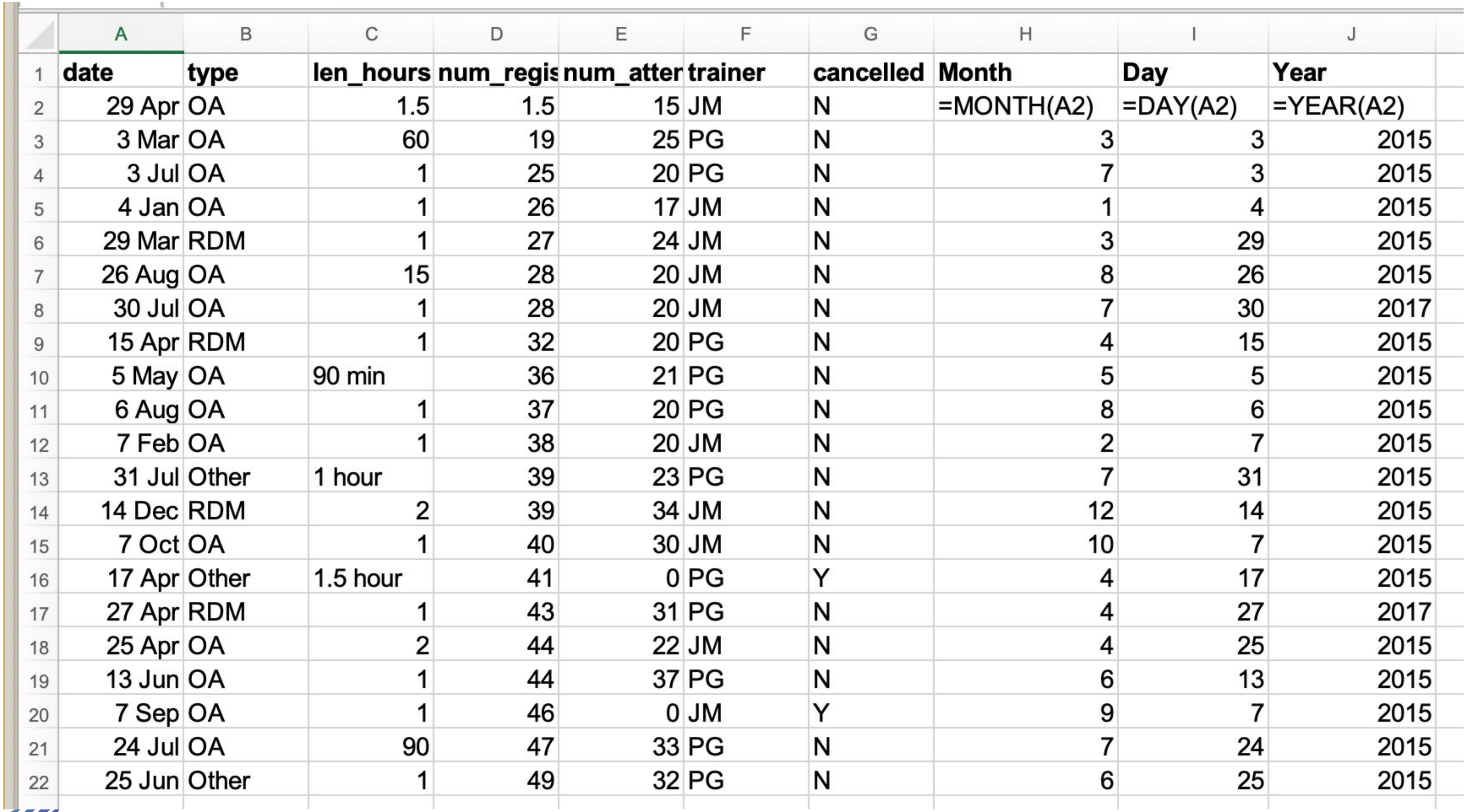
Date-related functions allow us to:

- convert date values from the stored numerical value to a readable display value
- make calculations between date values
- extract the date values so that they do not change as data is transformed or exchanged between new users and systems.

Please remember that functions that are valid for one spreadsheet program are usually guaranteed to be compatible only within the same family of products.

ACENET
accelerate discovery

Action of function	Excel	LibreOffice
Return the year number represented in the referenced cell value	`YEAR()`	`YEAR()`
Return the month number represented in the referenced date serial number	`MONTH()`	`MONTH()`
Return the day of the month represented in the referenced date serial number	`DAY()`	`DAY()`
Calculate and display a date based on supplied year, month, and day values	`DATE(Year, Month, Day)`	`DATE(Year; Month; Day)`
Return the serial number for date information supplied as a string	`DATEVALUE()`	`DATEVALUE("Text")`
Change display of a number by applying specified formatting	`TEXT(Value, "Formatting code to apply")`	`TEXT(Value; "Formatting to apply")`
Return the current system date	`NOW()`	`NOW()`



	A	B	C	D	E	F	G	H	I	J
1	**date**	**type**	**len_hours**	**num_regis**	**num_atter**	**trainer**	**cancelled**	**Month**	**Day**	**Year**
2	29 Apr	OA	1.5	1.5	15	JM	N	=MONTH(A2)	=DAY(A2)	=YEAR(A2)
3	3 Mar	OA	60	19	25	PG	N	3	3	2015
4	3 Jul	OA	1	25	20	PG	N	7	3	2015
5	4 Jan	OA	1	26	17	JM	N	1	4	2015
6	29 Mar	RDM	1	27	24	JM	N	3	29	2015
7	26 Aug	OA	15	28	20	JM	N	8	26	2015
8	30 Jul	OA	1	28	20	JM	N	7	30	2017
9	15 Apr	RDM	1	32	20	PG	N	4	15	2015
10	5 May	OA	90 min	36	21	PG	N	5	5	2015
11	6 Aug	OA	1	37	20	PG	N	8	6	2015
12	7 Feb	OA	1	38	20	JM	N	2	7	2015
13	31 Jul	Other	1 hour	39	23	PG	N	7	31	2015
14	14 Dec	RDM	2	39	34	JM	N	12	14	2015
15	7 Oct	OA	1	40	30	JM	N	10	7	2015
16	17 Apr	Other	1.5 hour	41	0	PG	Y	4	17	2015
17	27 Apr	RDM	1	43	31	PG	N	4	27	2017
18	25 Apr	OA	2	44	22	JM	N	4	25	2015
19	13 Jun	OA	1	44	37	PG	N	6	13	2015
20	7 Sep	OA	1	46	0	JM	Y	9	7	2015
21	24 Jul	OA	90	47	33	PG	N	7	24	2015
22	25 Jun	Other	1	49	32	PG	N	6	25	2015

Adding dates - using DATE() function

Adding years and months and days is slightly trickier because we need to make sure that we are adding the amount to the correct entity.

- First we extract the single entities (**day, month, or year**)
- We can then add **values** to do that
- Finally, the complete date string is reconstructed using the **DATE() function**.

Note: Time values raise similar challenges. Seconds can be directly added but to add hours and minutes you will need to make sure that quantities are added to the correct entities.

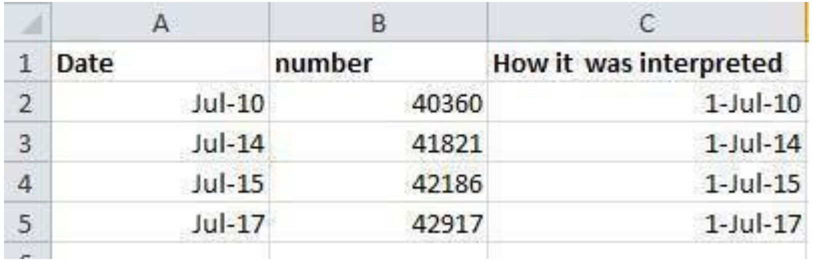
Advantages of alternative date formatting

The display ambiguities discussed can lead to unintended changes or unknown errors in your data. Exchanging data between applications or converting data into different formats can also create unexpected changes, and cause challenges for data interoperability, sharing & reuse, and long-term preservation.

Alternative date formats can help address these issues:

1. Storing dates as **YEAR, MONTH, DAY**
2. Storing dates as **YEAR, DAY-OF-YEAR**
3. Storing dates & times as **a single string**

YEAR, MONTH, DAY

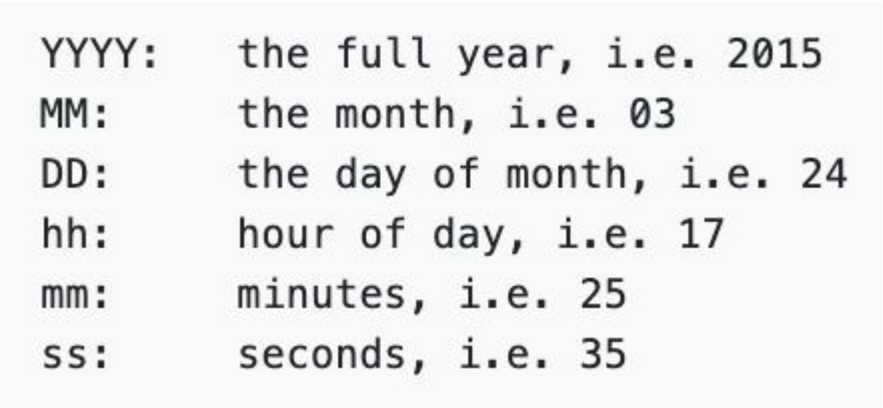


	A	B	C
1	Date	number	How it was interpreted
2	Jul-10	40360	1-Jul-10
3	Jul-14	41821	1-Jul-14
4	Jul-15	42186	1-Jul-15
5	Jul-17	42917	1-Jul-17

In dealing with dates in spreadsheets, we recommend separating date data into separate fields (**day, month, year**), which will eliminate any chance of ambiguity.

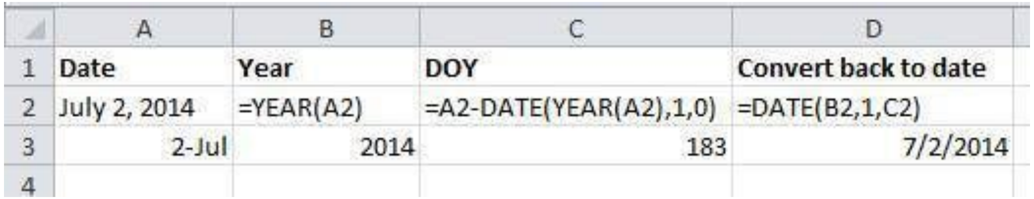
YYYYMMDDhhmmss format

March 24, 2015 17:25:35 ⟹ 20150324172535



```
YYYY:   the full year, i.e. 2015
MM:     the month, i.e. 03
DD:     the day of month, i.e. 24
hh:     hour of day, i.e. 17
mm:     minutes, i.e. 25
ss:     seconds, i.e. 35
```

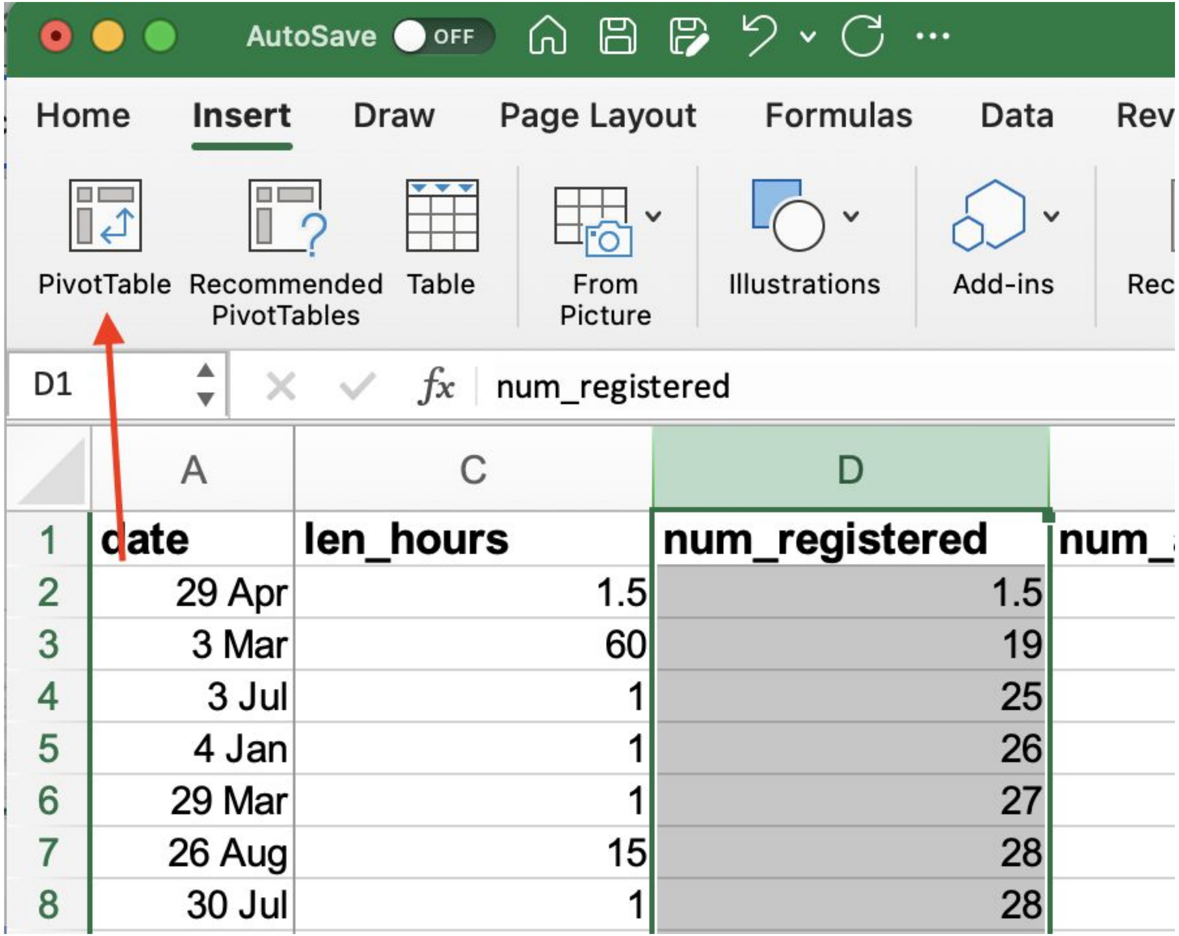
YEAR, DAY-OF-YEAR (DOY)



	A	B	C	D
1	Date	Year	DOY	Convert back to date
2	July 2, 2014	=YEAR(A2)	=A2-DATE(YEAR(A2),1,0)	=DATE(B2,1,C2)
3	2-Jul	2014	183	7/2/2014
4				

Pivot tables

A pivot chart is a graphical representation of data created from a pivot table. It allows users to visualize and analyze data in a more interactive and dynamic manner. Pivot charts are valuable for summarizing and presenting large datasets, making it easier to draw insights and trends from the data.



D1 num_registered

	A	C	D	
1	date	len_hours	num_registered	num_
2	29 Apr	1.5	1.5	
3	3 Mar	60	19	
4	3 Jul	1	25	
5	4 Jan	1	26	
6	29 Mar	1	27	
7	26 Aug	15	28	
8	30 Jul	1	28	

Where to find pivot tables in your program

Microsoft Excel: Under the “Insert” tab in the Excel ribbon

Google Sheets: Click on “Insert” >> Pivot Table

LibreOffice: Choose the Insert Pivot Table command from the main menu or click the from the Standard toolbar.

Apple Numbers: In the Numbers menu bar at the top of your screen, choose Organize > Create Pivot Table

Make sure your data is cleaned and errors/inconsistencies are removed before inserting a pivot table!

How to create a pivot table

1. Select the columns you want to highlight, or use the whole dataset
   a. Let’s highlight C, D, & E
2. Click on “Insert Pivot Table” or choose where to place it in the existing worksheet.
   a. Let’s choose somewhere near to the rest of the columns, starting at M:1
3. Set up the pivot table
   a. Drag the “category” field to the “Rows” area
   b. Drag the “quantity” fields to the “Values” area
4. Create the Pivot Chart:
   a. With the pivot table selected, go to the “Insert” tab in the Excel Ribbon.
   b. Click on the “PivotChart” button and choose the chart type you want to create (e.g., a column chart).
   c. A new chart will be generated next to the pivot table

How to create a pivot table

- Customize the Pivot Chart:
  a. Add chart titles, axis labels, and data labels to enhance the readability of the chart.
- Interact with the Pivot Chart:
  a. As the pivot table data changes, the pivot chart will automatically update to reflect the new data.
  b. You can also use the drop-down arrows on the pivot chart to filter and drill down into specific data subsets.

Pivot charts offer dynamic capabilities that make data exploration and analysis more interactive and effective.

Basic quality assurance & control

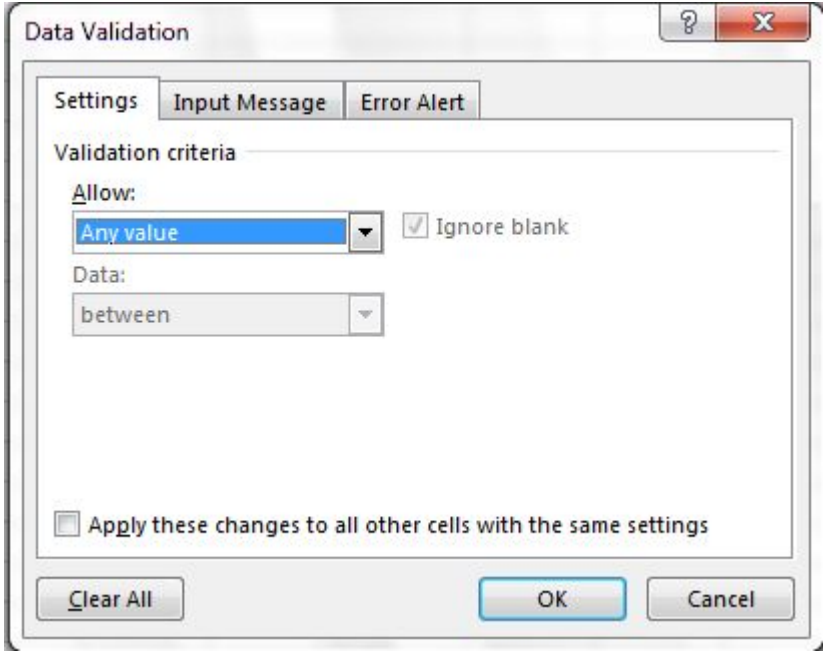
When you have a well-structured data table, you can use several simple techniques within your spreadsheet to ensure the data you enter is free of errors. These approaches include:

- techniques that are implemented prior to entering data **(quality assurance)**
- techniques that are used after entering data to check for errors **(quality control)**.

Quality assurance

Quality assurance stops bad data from ever being entered by checking to see if values are valid during data entry.

To control the kind of data entered into a spreadsheet we use **Data Validation** (Excel, Google Sheets) or **Validity** (LibreOffice Calc), to set the values that can be entered in each data column.



Data Validation

Validation criteria

Allow: Any value

☑ Ignore blank

Data: between

□ Apply these changes to all other cells with the same settings

Where to find data validation in your program

Microsoft Excel: Under the “Data” tab in the Excel ribbon

Google Sheets: Click on “Data” >> Data Validation

LibreOffice: Click on “Data” >> Validity

Apple Numbers: Numbers doesn’t have a data validation option. You can, however list all your values in column, select them, and set the Data Format (under Cell) to Pop-Up Menu. This will create a Pop-up Menu in each cell. You can copy-paste one anywhere you like.

Quality control

readme (README) files:

As you start manipulating your data files, create a readme document / text file to keep track of your files and document your manipulations so that they may be easily understood and replicated. Your readme file should document:

- all of the files in your data set (including documentation),
- describe their content and format
- lay out the organizing principles of folders and subfolders.
- for each of the separate files listed, document the manipulations or analyses that were carried out on those data.

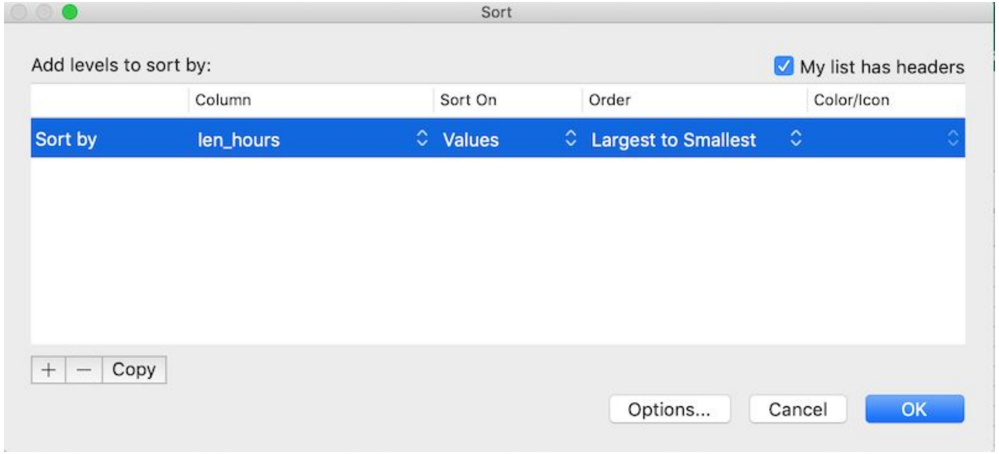
Cornell University’s Research Data Management Service Group provides detailed guidelines for how to write a good readMe file, along with an adaptable template.

Sorting

Bad values often sort to bottom or top of the column.

For example, if your data should be numeric, then alphabetical and null data will group at the ends of the sorted data. Sort your data by each field, one at a time. Scan through each column, but pay the most attention to the top and the bottom of a column. If your dataset is well-structured and does not contain formulas, sorting should never affect the integrity of your dataset.

Let’s try this in our Dates tab:



Sort

Add levels to sort by:

☑ My list has headers

	Column	Sort On	Order	Color/Icon
Sort by	len_hours	Values	Largest to Smallest	

support@ace-net.ca

Conditional formatting

- Conditional formatting basically can do something like **colour code your values by some criteria or from lowest to highest**.
- This makes it easy to scan your data for outliers.
- Use with caution! We can also do these checks in a programming language like Python or R, or in OpenRefine or SQL.

Let's try this in the Dates tab:

1. Make sure the `num_attended` column is highlighted.
2. Go to **Format** then **Conditional Formatting**, click +.
3. Apply any **2-Color Scale formatting rule**.
4. Now we can scan through and different colors will stand out. Do you notice any strange values?

Exporting data from spreadsheets

Exporting cleaned data as `.csv`

It is not a good idea to store the data you're going to work with for your analyses in Excel file formats (`.xls` or `.xlsx`). **Why?**

- Excel is a proprietary format - it is possible it may not exist in the future (think about zip drives/floppy disks!)
- Other spreadsheet software may not be able to open Excel format
- Different versions of Excel may handle data differently, leading to inconsistencies

ACENET
accelerate discovery

support@ace-net.ca

Storing data in a universal, open, static format

Try tab-delimited or CSV (more common). CSV files are plain text files where the columns are separated by commas, hence ‘comma separated variables.’

Advantages:

- We can open and read a CSV file using **almost any software**
- Data in a CSV can also be easily imported into other forms and environments, such as **SQLite and R**
- Offers **maximum portability and endurance**

Backwards compatibility: you can open CSVs in Excel!

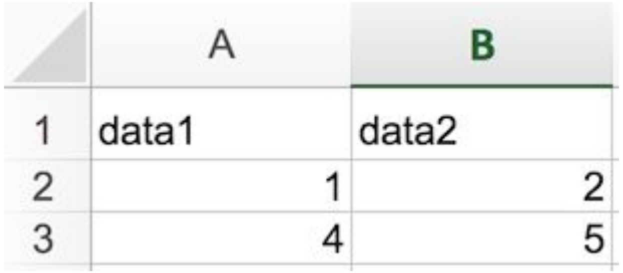
A note on cross-platform operability

By default, most coding and statistical environments expect UNIX-style line endings (ASCII LF character) as representing line breaks. However, Windows uses an alternate line ending signifier (ASCII CR LF characters) by default for legacy compatibility with Teletype-based systems..

When exporting to CSV using Excel, your data in text format will look like this:

```
data1,data2<CR><LF>1,2<CR><LF>4,5<CR><LF>
```

When opening your CSV file in Excel again, it will parse as follows:



	A	B
1	data1	data2
2	1	2
3	4	5

A note on Python & R

There are Python & R packages that can read xls files (as well as Google Sheets). It is even possible to access different worksheets in the xls documents.

But:

- this equates to replacing a (simple but manual) export to csv with additional complexity/dependencies in the data analysis Python code
- data formatting best practices STILL apply
- Is there really a good reason why csv (or similar) is not adequate?

Combining Excel with other tools

Data cleaning/data wrangling? Consider OpenRefine

- Transform data to make it more appropriate and valuable for a variety of purposes such as data analytics
- Easy to use – uses ‘clustering technology’: easy to find and correct spelling variations
- No need to learn programming language syntax or commands
- Can handle tens of thousands of rows

Advanced data analytics? Consider R or Python

- Both can handle much larger volumes of data, and therefore, more analysis
- **R** is designed to be reproducible and to create more detailed visualizations
- **Python** can replace mundane tasks with automation
  - Greater efficiency and scalability
  - Faster for automation and calculating complex equations, algorithms

support@ace-net.ca

Further resources

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

White, E., Baldridge, E., Brym, Z., Locey, K., McGlinn, D., & Supp, S. (2013). Nine simple ways to make it easier to (re)use your data. Ideas in Ecology and Evolution, 6(2). https://doi.org/10.4033/iee.2013.6b.6.f

Date & time sources

- The Earth Systems Research Lab provides this calendar that displays day of year information for any year you select: https://www.esrl.noaa.gov/gmd/grad/neubrew/Calendar.jsp.
- The U.S. National Snow and Ice Data Center provides a useful chart to calculate the day of year: https://nsidc.org/support/faq/day-year-doy-calendar.
- Microsoft Excel date and time functions reference.
- LibreOffice Date & Time Functions Reference v. 6.2.

ACENET

accelerate discovery

ace-net.ca

info@ace-net.ca

support@ace-net.ca

A **regional partner** of the

Digital Research
Alliance of Canada