
# Data Carpentry

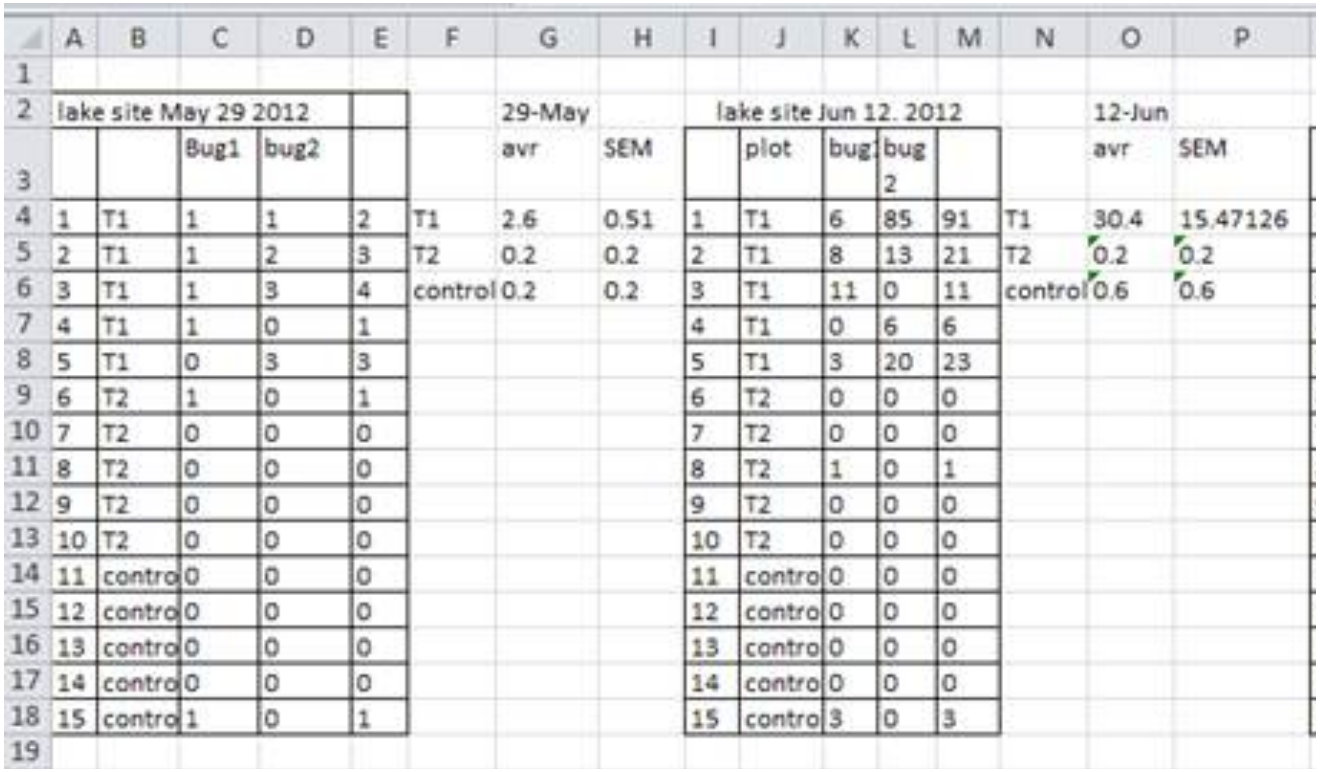
## Data Organization in Spreadsheets

# Introduction

# Introduction

Good data organization is the foundation of any research project. Much of your time as a researcher is spent in ‘data wrangling’.

# Data entry and organization is the start of any computational project



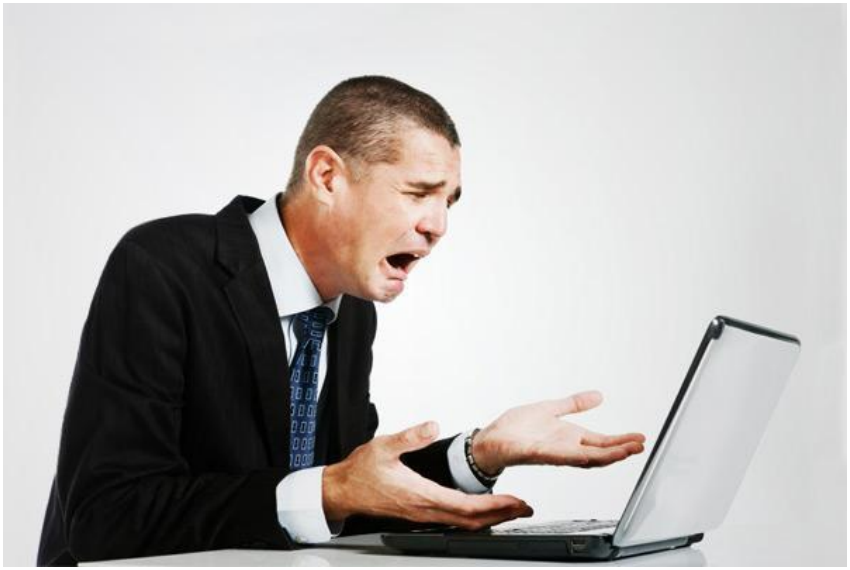
| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | | | | | | |
| 2 | lake site May 29 2012 | | | | | | 29-May | | lake site Jun 12. 2012 | | | | | | 12-Jun | |
| 3 | | | Bug1 | bug2 | | | avr | SEM | | plot | bug1 | bug 2 | | | avr | SEM |
| 4 | 1 | T1 | 1 | 1 | 2 | T1 | 2.6 | 0.51 | 1 | T1 | 6 | 85 | 91 | T1 | 30.4 | 15.47126 |
| 5 | 2 | T1 | 1 | 2 | 3 | T2 | 0.2 | 0.2 | 2 | T1 | 8 | 13 | 21 | T2 | 0.2 | 0.2 |
| 6 | 3 | T1 | 1 | 3 | 4 | control | 0.2 | 0.2 | 3 | T1 | 11 | 0 | 11 | control | 0.6 | 0.6 |
| 7 | 4 | T1 | 1 | 0 | 1 | | | | 4 | T1 | 0 | 6 | 6 | | | |
| 8 | 5 | T1 | 0 | 3 | 3 | | | | 5 | T1 | 3 | 20 | 23 | | | |
| 9 | 6 | T2 | 1 | 0 | 1 | | | | 6 | T2 | 0 | 0 | 0 | | | |
| 10 | 7 | T2 | 0 | 0 | 0 | | | | 7 | T2 | 0 | 0 | 0 | | | |
| 11 | 8 | T2 | 0 | 0 | 0 | | | | 8 | T2 | 1 | 0 | 1 | | | |
| 12 | 9 | T2 | 0 | 0 | 0 | | | | 9 | T2 | 0 | 0 | 0 | | | |
| 13 | 10 | T2 | 0 | 0 | 0 | | | | 10 | T2 | 0 | 0 | 0 | | | |
| 14 | 11 | contro | 0 | 0 | 0 | | | | 11 | contro | 0 | 0 | 0 | | | |
| 15 | 12 | contro | 0 | 0 | 0 | | | | 12 | contro | 0 | 0 | 0 | | | |
| 16 | 13 | contro | 0 | 0 | 0 | | | | 13 | contro | 0 | 0 | 0 | | | |
| 17 | 14 | contro | 0 | 0 | 0 | | | | 14 | contro | 0 | 0 | 0 | | | |
| 18 | 15 | contro | 1 | 0 | 1 | | | | 15 | contro | 3 | 0 | 3 | | | |
| 19 | | | | | | | | | | | | | | | | |

We often start by entering data into spreadsheets or getting tabular data from others or instruments

# How many people use spreadsheets in their work?

How many people have done something in spreadsheets that made them frustrated or sad?

# How many people have done something in spreadsheets that made them frustrated or sad?



We all use spreadsheets, but already know some of their limitations

# What this lesson teaches

Principles of data organization, cleaning, saving and sharing

We'll teach how to think about data organization and some practices for more effective data wrangling.

# What this lesson doesn't teach

- How to do analysis, plotting, statistics or write code in spreadsheets

# Why don’t we teach this?

This lesson is designed to help you move beyond spreadsheets to use other tools that allow you to work more effectively and reproducibly

- Data analysis in spreadsheets usually requires **a lot of manual work**. If you want to change a parameter or run an analyses with a new dataset, you usually have to redo everything by hand. (We do know that you can create macros, but see the next point.)

- It is also **difficult to track or reproduce statistical or plotting analyses** done in spreadsheet programs when you want to go back to your work or someone asks for details of your analysis.

# Lesson overview

- Principles of data organization in spreadsheets for computational work
- Guidelines for data cleaning
- Saving and sharing data

# Data organization

The tools we'll teach in this workshop will let you do more effective and reproducible work

# But it all starts with good data organization…

- We organize data like humans
- We need to organize data for computers

# Humans

**Plot: 2**

| Date collected | Species | Sex | Weight | Hindfoot_length |
|---|---|---|---|---|
| 1/8/14 | NN | F | | 7 cm |
| 1/8/14 | DM | M | 44 | 24 cm |
| 1/8/14 | DM | M | 38 | 23 cm |
| 1/8/14 | OL | F | | 232 cm |
| 1/8/14 | PE | M | 22 | 22 cm |
| 1/8/14 | DM | M | 38 | 26 cm |
| 1/8/14 | DM | M | 48 | 8 cm |
| 1/8/14 | DM | M | 43 | |
| 1/8/14 | DM | F | 35 | |
| 1/8/14 | DM | M | 43 | 182 cm |
| 1/8/14 | DM | F | 37 | 29 cm |
| 1/8/14 | PF | F | 7 | 115 cm |
| 1/8/14 | DM | M | 45 | 190 cm |
| 1/8/14 | OT | F | | 122 cm |
| 1/8/14 | DS | M | 157 | 133 cm |
| 1/8/14 | OX | F | | 107 cm |
| 2/18/14 | NN | M | 218 | 99 cm |
| 2/18/14 | PF | F | 7 | 84 cm |
| 2/18/14 | DM | M | 52 | 88 cm |

scale not calibrated

# Computers

| Day | Month | Year | Plot | Species | Sex | Weight | Hindfoot | Calibrated |
|---|---|---|---|---|---|---|---|---|
| 8 | 1 | 2014 | 2 | NN | F | NA | 7 | Y |
| 8 | 1 | 2014 | 2 | DM | M | 44 | 24 | Y |
| 8 | 1 | 2014 | 2 | DM | M | 38 | 23 | Y |
| 8 | 1 | 2014 | 2 | OL | F | NA | 232 | Y |
| 8 | 1 | 2014 | 2 | PE | M | 22 | 22 | Y |
| 8 | 1 | 2014 | 2 | DM | M | 38 | 26 | Y |
| 8 | 1 | 2014 | 2 | DM | M | 48 | 8 | Y |
| 8 | 1 | 2014 | 2 | DM | M | 43 | NA | Y |
| 8 | 1 | 2014 | 2 | DM | F | 35 | NA | Y |
| 8 | 1 | 2014 | 2 | DM | M | 43 | 182 | Y |
| 8 | 1 | 2014 | 2 | DM | F | 37 | 29 | Y |
| 8 | 1 | 2014 | 2 | PF | F | 7 | 115 | Y |
| 8 | 1 | 2014 | 2 | DM | M | 45 | 190 | Y |
| 8 | 1 | 2014 | 2 | OT | F | NA | 122 | Y |
| 8 | 1 | 2014 | 2 | DS | M | 157 | 133 | N |
| 8 | 1 | 2014 | 2 | OX | F | NA | 107 | Y |
| 18 | 2 | 2014 | 2 | NN | M | 218 | 99 | N |
| 18 | 2 | 2014 | 2 | PF | F | 7 | 84 | Y |
| 18 | 2 | 2014 | 2 | DM | M | 52 | 88 | Y |

# Rules for structuring data in spreadsheets

- Put all your **variables in columns**
  the thing you're measuring, like 'weight' or 'temperature'.
- Put each **observation in its own row**.
- **Don't combine multiple pieces of information in one cell**.
  Sometimes it just seems like one thing, but think if that's the only way you'll want to be able to use or sort that data.
- Meaningful **column names** with no spaces or special characters

# Exercise

- Download our messy spreadsheet
- Pair with a person next you and look for the things that are wrong with the spreadsheet.

# Exercise

- Download our messy spreadsheet
- Pair with a person next you and look for the things that are wrong with the spreadsheet.

- Discuss as a group: What problems did people find? See more information on the ‘common-mistakes’ page.

# Dates as data

Instead of using one column for dates, treat it as three separate sets of information

- Day, Month and Year

No date formatting issues and easier access to elements of the date

| Date collected |
|---|
| 1/8/14 |
| 1/8/14 |
| 1/8/14 |
| 2/18/14 |
| 2/18/14 |
| 2/18/14 |

| Day | Month | Year |
|---|---|---|
| 8 | 1 | =year(D5) |
| 8 | 1 | 2014 |
| 8 | 1 | 2014 |
| 18 | 2 | 2014 |
| 18 | 2 | 2014 |
| 18 | 2 | 2014 |

# Guidelines for data cleaning

# Humans → Computers

**Plot: 2**

| Date collected | Species | Sex | Weight | Hindfoot_length |
|---|---|---|---|---|
| 1/8/14 | NN | F | | 7 cm |
| 1/8/14 | DM | M | 44 | 24 cm |
| 1/8/14 | DM | M | 38 | 23 cm |
| 1/8/14 | OL | F | | 232 cm |
| 1/8/14 | PE | M | 22 | 22 cm |
| 1/8/14 | DM | M | 38 | 26 cm |
| 1/8/14 | DM | M | 48 | 8 cm |
| 1/8/14 | DM | M | 43 | |
| 1/8/14 | DM | F | 35 | |
| 1/8/14 | DM | M | 43 | 182 cm |
| 1/8/14 | DM | F | 37 | 29 cm |
| 1/8/14 | PF | F | 7 | 115 cm |
| 1/8/14 | DM | M | 45 | 190 cm |
| 1/8/14 | OT | F | | 122 cm |
| 1/8/14 | DS | M | 157 | 133 cm |
| 1/8/14 | OX | F | | 107 cm |
| 2/18/14 | NN | M | 218 | 99 cm |
| 2/18/14 | PF | F | 7 | 84 cm |
| 2/18/14 | DM | M | 52 | 88 cm |

scale not calibrated

| Day | Month | Year | Plot | Species | Sex | Weight | Hindfoot | Calibrated |
|---|---|---|---|---|---|---|---|---|
| 8 | 1 | 2014 | 2 | NN | F | NA | 7 | Y |
| 8 | 1 | 2014 | 2 | DM | M | 44 | 24 | Y |
| 8 | 1 | 2014 | 2 | DM | M | 38 | 23 | Y |
| 8 | 1 | 2014 | 2 | OL | F | NA | 232 | Y |
| 8 | 1 | 2014 | 2 | PE | M | 22 | 22 | Y |
| 8 | 1 | 2014 | 2 | DM | M | 38 | 26 | Y |
| 8 | 1 | 2014 | 2 | DM | M | 48 | 8 | Y |
| 8 | 1 | 2014 | 2 | DM | M | 43 | NA | Y |
| 8 | 1 | 2014 | 2 | DM | F | 35 | NA | Y |
| 8 | 1 | 2014 | 2 | DM | M | 43 | 182 | Y |
| 8 | 1 | 2014 | 2 | DM | F | 37 | 29 | Y |
| 8 | 1 | 2014 | 2 | PF | F | 7 | 115 | Y |
| 8 | 1 | 2014 | 2 | DM | M | 45 | 190 | Y |
| 8 | 1 | 2014 | 2 | OT | F | NA | 122 | Y |
| 8 | 1 | 2014 | 2 | DS | M | 157 | 133 | N |
| 8 | 1 | 2014 | 2 | OX | F | NA | 107 | Y |
| 18 | 2 | 2014 | 2 | NN | M | 218 | 99 | N |
| 18 | 2 | 2014 | 2 | PF | F | 7 | 84 | Y |
| 18 | 2 | 2014 | 2 | DM | M | 52 | 88 | Y |

# Guidelines for data cleaning

- **Leave the raw data raw**

  If you don't have an original of the data you can never replicate your analyses to know they're correct

- **Keep track of the cleaning steps**

  Take notes, or have a script

- Save the cleaned data in a **text based format** like comma separated values (CSV)

  This ensures that anyone can use the data, and is the format required by most data repositories.

# Exercise

- With the messy spreadsheet, start to clean up one section of the data following the data cleaning guidelines.

# Exercise

- With the messy spreadsheet, start to clean up one section of the data following the data cleaning guidelines.
- Trade notes with the person next to you. Imagine them as the you of 6 months from now. Can they figure out what you did? (Past you is terrible at answering email)

# Data saving and sharing

# Guidelines for data saving and sharing

- **Export your data** as comma or tab separated values (CSV or TSV). This format is the one that can be read by other programs and does not rely on particular software.
- Raw data, and anything you don't want to have to redo, should be **backed up**.
- Give data **meaningful file names**. Using dates in the names helps for making them unique and findable.
- Data about the data (**metadata**) also should be tracked and saved.
- Data should be **available to someone other than just you**. (Pass the 'decide to bike around the world' test)

# Exercise

- Think about the data you have right now. Is it backed up? If so, how? Do people other than you have access to the data and could they understand what it is?

# Exercise

- Think about the data you have right now. Is it backed up? If so, how? Do people other than you have access to the data and could they understand what it is?

- Discuss as a group: What were some options you had for saving and sharing data? (There's no one solution and it depends on data type and size)

# Questions or comments?