# Data Wrangling: Assignment 1

Due Monday, January 22 at midnight MST

## Purpose:

Describe and practice good data entry techniques in spreadsheets.

## Task:

Explain issues in a messy dataset, create a tidy version of the messy dataset, and create a spreadsheet with effective data validation.

## Criteria for Success:

To successfully complete this assignment, you should turn in:

1. A document with written answers (Ex. 1, Q1-Q3).
2. A .csv file that demonstrates best-practices in creating tidy data that is human- and computer-readable (Ex. 1, Q4).
3. A .xlsx (not .csv in this case) that has the correct data validation criteria (Ex. 2).

## Exercises:

1. Improving Messy Data (50 pts)

A lot of real data isn’t very tidy, mostly because most scientists aren’t taught about how to structure their data in a way that is easy to analyze.

[Download an untidy version](https://datacarpentry.org/semester-biology/data/untidy-portal-data.xlsx) of some of the Portal Project data, which includes information on the site, date, species identification, weight, and sampling plot (within the site) for some small mammals.

Think about what could be improved about this data and write down answers to the following questions:

* 1. Describe five things about this data that are not tidy and how you could fix each of those issues.
     1. In the Plot 1 table there are numbers and units (“g”) in the weight column.
     2. There are highlighted cells in the Plot 2 table. This information about incorrect calibration won’t be carried over into whatever programming language/IDE you choose to analyze this data with.
     3. There are spaces in column names/the spacing/naming conventions aren’t consistent across the file.
     4. The dates in the “Date collected” column for each table are not in the same format.
     5. Plot 4 needs to be pivoted so that the species names all fall into a column called “species” or something similar. Generally, having 4 tables for 4 different plots/measurements in one file is not a good way to do things. There should be a column for “plot” with a number designating which plot each observation falls into.
  2. Could this data easily be imported into a programming language or a database in its current form? Why or why not?
     1. No! Given that the data is not a rectangle it would be a nightmare to import this into a database or language. The computer would not know what to do with the spaces, inconsistent naming conventions, etc.
  3. Do you think it’s a good idea to enter the data like this and clean it up later, or to have a good data structure for analysis by the time data is being entered? Why?
     1. I think it is better to have a good data structure set up by the time data is entered so you can get right into analysis and checking out what your data says. Additionally, it is easier to rearrange data or add columns for new variables of interest if your data is already set up in a tidy way.
  4. Based on the best practices you’ve learned, create a spreadsheet that presents the data in a tidy format and is both human- and computer-readable. You don’t need to add in all of the data from the tables, but you should have (good) column names and ~5 rows of example data. Save the data file as a CSV file.

1. Data entry validation in Excel (50 pts)

Create a spreadsheet in Excel for data entry. It should have five columns: Date, Site, Species, Mass, and Length.

Set the following data validation criteria to prevent invalid data from getting entered:

* 1. The Date column should be set so that it doesn’t convert dates to other formats.
  2. Use data validation so that Site can only be one of the following A1, A2, B1, B2. Set the error message on this validation criteria to provide information on what the valid values are.
  3. Use data validation so that Species can only be one of the following Dipodomys spectabilis, Dipodomys ordii, Dipodomys merriami. Set the error message on this validation criteria to provide information on what the valid values are.
  4. Use data validation so that Mass can only be a decimal greater than or equal to zero but less than or equal to 500. Set the error message on this validation criteria to provide information on what the valid values are.
  5. Length should be an integer (i.e., a whole number) between 1 and 10. Set the error message on this validation criteria to provide information on what the valid values are.

Check that the validation rules and data formatting are working, but do not include any entered data in the final file.

Save this file as data\_entry\_form.xlsx.

In this case, I don’t want you to save the file as a .csv because I want to check the data validation measures specifically.