|  |  |
| --- | --- |
| **“Cleaning” Biodiversity Data in Excel** |  |

**Objectives**

Upon completion of this module, each student should be able to:

1. Access biodiversity data from open sources.
2. Use descriptive, retrievable, and consistent file names to manage datasets.
3. Identify common problems with digital datasets
4. Rectify common problems with digital datasets
5. Apply disciplinary knowledge for smart data cleaning
6. Explain the importance of reproducible data and cleaning steps
7. Document data cleaning steps to provide reproducibility.

**Introduction**

The biodiversity sciences have experienced a rapid mobilization of data that has increased our capacity to investigate large-scale issues. The volume and variety of data being generated and the increased accessibility of data for aggregation are driving the need for new skill sets. One of these skills is “data cleaning”. Data cleaning involves identifying errors or inconsistencies in the data and then correcting these issues. To err is human, so any dataset may have flaws, but the increase in data aggregation from multiple sources increases the likelihood of these errors and inconsistencies in the way the data are reported.

Aggregating data from potentially hundreds of sources requires coming to an agreement on the way the data will be described and recorded. These agreements are called data standards. In order to share, exchange, and understand data, we must standardize data formats and agree on the terms that will be used to identify different types of data. Biodiversity data from aggregators like iDigBio or GBIF conform to a specific set of rules called DarwinCore. At the most basic level, DarwinCore is a standard set of named data fields that should be filled in as completely as possible, and in standardized formats, for each specimen record.

Data aggregators are constantly working to improve their ability to seamlessly integrate multiple data sources, but it is still your responsibility as a data user to examine your data before moving forward with your analysis. Some of the most common problems you may find in datasets are:

* Duplicate records [the same specimen is represented twice in your dataset]
* Missing data [not all data cells are filled for each specimen record]
* Inaccurate data [e.g., errors in data entry; errors in species identifications]
* Inconsistent data [e.g., masses reported in different units; species name changes]
* Outliers [e.g., location points that seem to lie outside the possible species range]

As you make decisions about how you will address different errors and inconsistencies in a dataset, it is crucial that you document the changes that you make. This documentation is necessary so that you, or another researcher, could repeat your processing and analysis. In science, we call this “reproducibility”? According to a U.S. National Science Foundation (NSF) subcommittee on replicability in science, “reproducibility refers to the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results…. Reproducibility is a minimum necessary condition for a finding to be believable and informative” (Bollen et. al, 2015).

Reproducible science is also often linked with the idea of research transparency. If each step in the process is well documented and shared it means that people outside of the original research team whether that be other scientists, policy makers, the press, or the general public, are better able to assess of the validity of a study. Creating a reproducible study means accurately and clearly designating each step you took to collect, clean, and analyze your data. When cleaning data it means keeping an accurate record of which data points you discarded or modified and why.

In this module, you will work through several of these issues while cleaning a data set provided by your instructor that includes some known errors. Then, you will download and clean your own data set and use it to investigate a species of your choosing.

**Note**: While these instructions provide directions for steps to take for data cleaning using Excel, different operating systems and software versions may have different interfaces for accessing the functions needed for data cleaning. You may need to use spend some time looking around your version to find some of the commands. You can use the “Help” feature in Excel to locate the commands, look online for tutorials, or ask your classmates or instructor for help.

Here is a preview of our Data Cleaning Workflow:

1. Save an informatively labeled raw data file.
2. Create an informatively labeled working copy of your data for analysis.
3. Import the data file into your working platform with correct formatting.
4. Remove duplicate records, if present.
5. Remove records without the data of interest.
6. Explore missing data and make decisions about which records to keep.
7. Standardize formats and units.
8. Explore possible inaccuracies and make decisions about which records to keep.
9. Explore possible outliers and make decisions about which records to keep.

**Activity 1: Prepare Your Working Data File**

Before making any modifications to a dataset, you must first create a duplicate copy of your data file. One of the copies will remain unmodified while you use the other to do your cleaning and analysis. Choose your file names carefully so that they are descriptive and consistent.

*Procedure*: *Your instructor will tell you where you can access the raw polar bear data file.*

1. Create a folder in which you will keep all files related to the module. Give the folder a descriptive title that includes the name of this course and something describing this activity. For example, “Data Cleaning Module”.
2. Download the data file, “*Ursus arctos* Raw Data” and save it in the folder.
3. Create a duplicate copy of the file in the same folder. Give the data file a different (still descriptive) title, for example “*Ursus arctos* Cleaned Data” … or “BIO 212 *Ursus arctos* Working Data”.
4. Now you are ready to begin cleaning.

**Activity 2: Remove Duplicate Records**

There are several programs that you could use to analyze data. In this activity, we will be using Microsoft Excel to clean your data. Some screen shots will be provided to help you locate different commands and tools; however, you may have a different version of Excel or be working in a different operating system. If you have trouble locating the commands, may find it necessary to use the program’s help feature, ask an instructor or classmate, or search online for assistance.

1. Open the file in excel.
2. Add a new sheet and name it “Remove Duplicates”.
3. Copy and paste the raw data into the new sheet.
4. In the “Data” tab, click on “Remove Duplicates”



1. In the “Remove Duplicates” window, select only the column that includes the data field GUID. This is a unique identifier. In the sample data set, this is Column B.



1. An alert window will pop up and tell you if any records have been removed and how many remain.

*Were any records removed from your data set? If so, how many were removed and how many records remain?*

**Activity 3: Remove “Incomplete” Records**

Missing data may be the easiest error to identify from a cursory inspection of your data file. An empty cell in your spreadsheet represents missing data. However, different collectors may have denoted missing data differently. Be aware of common markers that data collectors may use to denote missing data that may interfere with analysis. These include NA, X, and 0.

Almost all specimen records have some missing data. In order to determine which records to retain for your analyses, you must know what your crucial variables of interest are. You may choose to retain a specimen record that contains at least one of your variables of interest, or you may decide that the record is only usable if it contains all of the variables you are investigating. The most important thing to remember is to document all of the decisions you make by keeping a copy of the spreadsheet at each stage of cleaning.

For the purposes of this activity, imagine that we are investigating the relationship between gender and weight for Brown Bear specimens in natural history collections.

*Procedure*

1. Insert a new sheet and copy all of the data from the “Remove Duplicates” sheet into the new sheet. Name the new sheet “Remove No Weight”.
2. In the “Data” tab, click on “Sort”.



1. Sort the spreadsheet using the column containing your variable of interest. For this activity, begin by sorting by “Weight”. Excel won’t recognize these as numbers because of the unit labels included in some of the cells, so sort in “A to Z” order.



1. Scroll down until you reach records that do not have any data in the Weight column. Select all of the records that do not have weight data and delete those records/rows.

*How many records were removed and how many records remain in your data set?*

1. Insert a new sheet and name it “Remove No Sex”. Copy the data from “Remove No Weight” into the new sheet. Sort by the “Sex” column and remove any records that do not have the sex of the specimen recorded.

*How many records were removed and how many records remain in your data set?*

*What did you decide to do with the specimen of “unknown” sex? Will this record be useful for your analysis?*

**Activity 4: Standardize Formatting of All Data**

In order to share, exchange, and understand data, we must standardize the format as well as the meaning of some key terms. Data standards are the rules for the formats in which data are recorded. Biodiversity data from aggregators like iDigBio or GBIF conform to a set of data standards called DarwinCore. Standards help create a shared understanding of how data are represented across providers. Even with standards in place, there are still inconsistencies found in biodiversity data that need to be resolved before continuing with analysis.

Fields such as date, have clear guidelines in DarwinCore but not all data providers adhere to them yet. Other fields such as sex don’t have a standardized vocabulary yet, so data providers vary greatly in how they report information on the sex of an organism if they report it at all. It is important to determine which fields are the crucial variables in your analysis and then make sure they are all formatted for consistency --- or standardized, before you continue on with your analysis.

*Procedure*

1. Insert a new sheet and name it “Standardize Weight Units”. Copy the data from “Remove No Sex” into the new sheet.
2. Sort by Weight.
3. Insert a new column to the right of the “Weight” column. Give the column the header “Weight Units”.
4. Enter the abbreviation for the units used in the Weight column. For example, if the weight listed is 1000lb, enter “lb” in the “Weight Units” column. This is just to keep track of the original units while we make conversions.
5. Select the “Weight” column and use the “advanced find and replace” tool to eliminate (e.g., replace lb with nothing) all of the units. Repeat for each different unit. Scan to make sure only numbers remain in the column.

|  |  |
| --- | --- |
| A screenshot of a cell phone  Description automatically generated | A screenshot of a cell phone  Description automatically generated |

1. Insert a new column to the right of the Weight Units column. Give the column the header “Weight (g)”.
2. Sort by “Weight Units”.
3. Enter formulas in this column to convert the weights in the “Weight” column into grams (g). For example, if the original weight was recorded in pounds (lb) enter a formula to multiply the original weight by 453.6. For those weights that were already in grams, simply re-enter the original weight.
4. Now, all of these weights can be compared because the units have been standardized.

**Activity 5: Make Decisions about Outliers**

The last data error types we will address are inaccurate data or outliers. Dealing with these error types requires the highest amount of content knowledge about the your organisms. An outlier is a data point that differs significantly from other observations. Is the data point a mistake, or could it represent natural variation? If it is a mistake you should discard the data point, but if it represents natural variation you may want to retain it. These are all decisions that a researcher must make before moving forward with analysis. Whatever you decide to do it is important that you keep a record of all of the steps, actions, and decisions you made while preparing your data for analysis so that other scientists may follow in your footsteps and repeat your analysis is needed.

*Procedure*

1. Insert a new column to the right of the Weight column. Give the column the header “Remove Weight Outliers”.
2. Sort the spreadsheet largest to smallest by “Weight (g)”
3. Scan down the column of weights. At any point does there seem to be an extreme break in the data (a very large difference between two consecutive weights)?
4. Some of these extreme records are labeled “immature”. You should remove these records if you are interested in comparing the weights of *adult* male and female brown bears.
5. There are several other records that are definitely outside of the reasonable range for an adult brown bear. You should remove these records.
6. There will still be a few records where you’re going to need to use your judgement. Make a decision about where you are placing the “cut-off” and remove any outlier points beyond this range.

*How many records were removed and how many records remain in your data set?*

While location isn’t relevant to our gender vs. weight question, it is a data type that often requires examination for outliers or errors. A quick way to visualize these location data in excel to look for outliers is to make a scatterplot of latitude vs longitude.

*Procedure*

1. Insert a new column to the right of the Remove Weight Outliers column. Give the column the header “Remove Location Outliers”.
2. Select the “Dec\_Lat” and “Dec\_Long” columns. In the “Insert” tab, select the scatterplot.



1. View the scatterplot. If any of the points are far removed from the other points, you may want to remove the point from the data set. Again, you’ll have to use your judgement and document whatever decision you make.

|  |  |
| --- | --- |
| No Outliers or Obvious Errors | Probable Outlier or Entry Error |

**Activity 6: Analyze your Data**

Now that you have removed duplicate and incomplete records, standardized data formats, and identified and removed outliers, your data set is “clean”. You have documented each step in the process with a saved dataset and recorded how many records have been removed and why they were removed. Your data are now ready to be used for your analyses.

Data analysis goes beyond the scope of this module. However, in case you’re interested, the relationship between sex and weight of the *Ursus arctos* specimens represented in digitized natural history collections, based on our data set, is shown below.



This is just one example of the types of analyses you can do with digitized natural history collections data. When working with digitized natural history collection data, always remember that unless specimens were collected with a systematic sampling protocol designed to investigate this question, we must be cautious in the interpretation of our results.

**References**

K. Bollen, J. T. Cacioppo, R. Kaplan, J. Krosnick, J. L. Olds, Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science (National Science Foundation, Arlington, VA, 2015).

<https://www.usgs.gov/products/data-and-tools/data-management/data-standards>

<https://datacarpentry.org/OpenRefine-ecology-lesson/>

<https://nces.ed.gov/pubs2007/curriculum/ls_dataerrors.asp>