## 

|  |  |
| --- | --- |
| **Building Biodiversity Datasets** | European frog-bit |

## Objectives

Upon completion of this module, students will be able to:

* Explain the role of data in science
* Apply vocabulary and concepts related to datasets
* Translate field sampling protocols to data collection methods
* Build a dataset template using appropriate structure for project
* Identify tidy data and create tidy data templates

**Introduction**

There is growing concern as human-influenced ecosystem changes such as agriculture, urban development, and climate change threaten global **biodiversity** (the diversity of biological organisms on Earth). In a world more interconnected than ever, organisms are frequently transported to regions they are not native and they would not naturally grow or live. If a nonnative organism causes negative economic or biological impacts in the new region, it is considered an **invasive species**. Invasive species can cause significant harm to native organisms, communities, and ecosystems as they outcompete native species for resources such as food and space or negatively impact abiotic factors such as water quality, nutrient availability, and landscapes. Economic losses and conservation efforts to manage biological invasions have cost nearly 1.3 trillion US dollars in the past five decades (Diagne 2021).

European frog-bit (EFB) is an invasive aquatic plant in the Great Lakes region of North America. It can be recognized by floating heart-shaped leaves and white flowers. It was intentionally introduced as an ornamental aquatic plant in Ottawa, Ontario in 1932 and has since spread throughout Ontario, Quebec, and the northeastern and midwestern United States (Minshall 1940, Dore 1968). In Michigan, EFB inhabits Great Lakes coastal wetlands and inland waterbodies such as lakes, rivers and streams, and wetlands. Saginaw Bay, Michigan is an area of concern due to heavy infestation of EFB since 2013. Conservation professionals believe it may be negatively impacting local species and environmental conditions, but so far research is limited.



Michigan distribution of EFB

**Saginaw Bay**

A picture containing outdoor, grass, tree, plant

Description automatically generated

You are interested in sampling aquatic plants in Saginaw Bay to determine whether EFB is associated with a decrease in native diversity, so that researchers and managers can understand its effects and respond accordingly. You have already determined the materials and methods you will need to use to answer your question, but there is one more very important step before you are ready to begin

sampling: prepare for data collection. Why is this such a crucial step? Data are fundamental in the study and preservation of biodiversity, as they allow us to inventory biological organisms or systems, track changes over time, and combine efforts from multiple researchers toward a common goal. By recording what we observe, we are making sure others can use the same information to verify or refute our conclusions. Without a solid plan for collecting, organizing, and storing our data, we risk losing the context or losing the data altogether.

**What exactly are data?** Data are the units of information. Data provide context for information and records of knowledge. Much of what you know about science comes from **data** that were collected, put in the context of a particular question or topic (becoming **information**), and then being disseminated to other members of the scientific community (becoming **knowledge**; Brackett 2013). Data take many forms, including lists, tables, or even notes in a notebook. One kind of data that you will encounter frequently are tabular datasets, which are groups of columns and rows with values representing each combination. In most cases, columns should contain the types of information collected (**fields**) and rows should contain **observations**.

**Data vs. Datum**

The word “**data**” is a plural noun used for multiple pieces of information. The word “**datum**” is the singular form of the noun used for a single piece of information. For example, one temperature measurement produces one datum, and temperature measurements at several points over time produce many data. You will almost always deal with many data, not just one individual datum, so “**data**” is the more commonly used form of the word.

*Columns store fields, or the types of information collected*

*Rows store individual observations*

|  |  |  |  |
| --- | --- | --- | --- |
| **Restaurant name** | **City** | **State** | **Year established** |
| Cochon | New Orleans | Louisiana | 2006 |
| Waiahole Poi Factory | Kaneohe | Hawaii | 2009 |
| Indian Pueblo Kitchen | Albuquerque | New Mexico | 1976 |
| Turtle Tower | San Francisco | California | 2000 |

Before you begin sampling in Saginaw Bay, you will need to create a data collection template that includes spaces for all of the data you need to collect. Listing all the fields you want to collect in a table format ensures you don’t forget what you need and allows you to easily see mistakes, missing data, or values that don’t line up with what you expect. Thinking about your potential data and planning for your analysis before conducting your fieldwork help to assure you are prepared to be efficient in the field, data are not lost, and the data you do gather can be interpreted in context and relevant analyses. Additionally, using a template makes sharing and storing data simpler and less prone to errors compared to transcribing paragraphs of notes from a notebook (which will still require a template).

**Activity I: Set up Your Workspace**

The first step in any project should be setting up your digital **workspace**, which is where you’ll store all your files. In this case, your workspace will be a single folder on your computer.

*Procedure*

1. Create a new folder anywhere on your computer. You will want to choose a location that is easily accessible where the folder will not get lost, so your computer’s Desktop and Documents may be good locations. Name the folder “BLUE\_BuildingDatasets\_yyyymmdd.” If you begin this module on August 1, 2021, your folder will be named “BLUE\_BuildingDatasets\_20210801.”

Why did we name the folder this way? Here are some important concepts to consider when choosing file and folder names. Names should:

* Be descriptive. Avoid very long file names and instead choose the minimum amount of information needed to remind you what is in the file or folder.
* Use capitalization and underscores to separate parts of the name.
* Avoid special characters such as !@$<>?\*#/”&. Most special characters are not machine-readable, meaning you will run into problems using these file names in a program like R or Python.
* Include dates in ISO 8601 format (yyyymmdd).
* **Shape

  Description automatically generated**Be consistent within a project. Ideally, multiple people should be able to use the same workspace with everyone knowing what all names mean and how to create their own names to match.

1. Save this Word document (this module) in your folder. Give it a descriptive name that will make sense to you later on and follows the guidelines above.

**Activity II: Create a Data Template from the Sampling Protocol**

The next step before you are ready for your plant sampling is creating a template for data collection.

*Procedure*

1. Review the sampling protocol below.
2. In the space below, make a list of all the data parameters that will be collected. For example, Step 5 of the protocol instructs you to determine whether EFB is flowering, so that is one parameter that will be collected. (*Hint: Not all variables that should be recorded are explicitly stated in sampling protocols. For example, you should always record the date and name of collector(s), but this is not always indicated.)*

**Protocol for Sampling Plant Diversity in Saginaw Bay**

Sampling will occur for a total of 24 days from June to August.

There are 96 plots already set up for sampling, and you will visit each plot once.

1. When you arrive at the sampling plot, use a GPS device to record the latitude and longitude.

2. Record the air temperature and humidity.

3. Record the depth of the water at the center of the plot using a standard meter stick.

4. Note whether European frog-bit is present or absent in the plot. If present, record the estimated percentage of the plot that is covered, as aerial percent cover (<https://cnhp.colostate.edu/cnhpblog/2009/11/24/field-techniques-percent-cover-estimation/>).

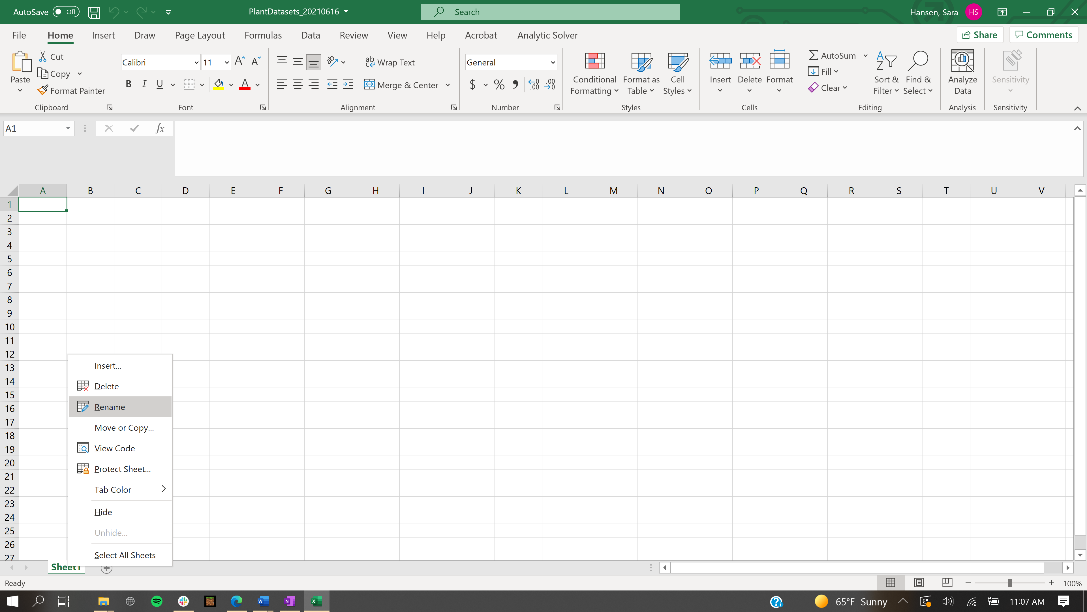
5. Examine European frog-bit for flowers.

**6. Record the estimated aerial percent cover each plant species. Because aquatic plants often overlap, total percent cover may exceed 100%.

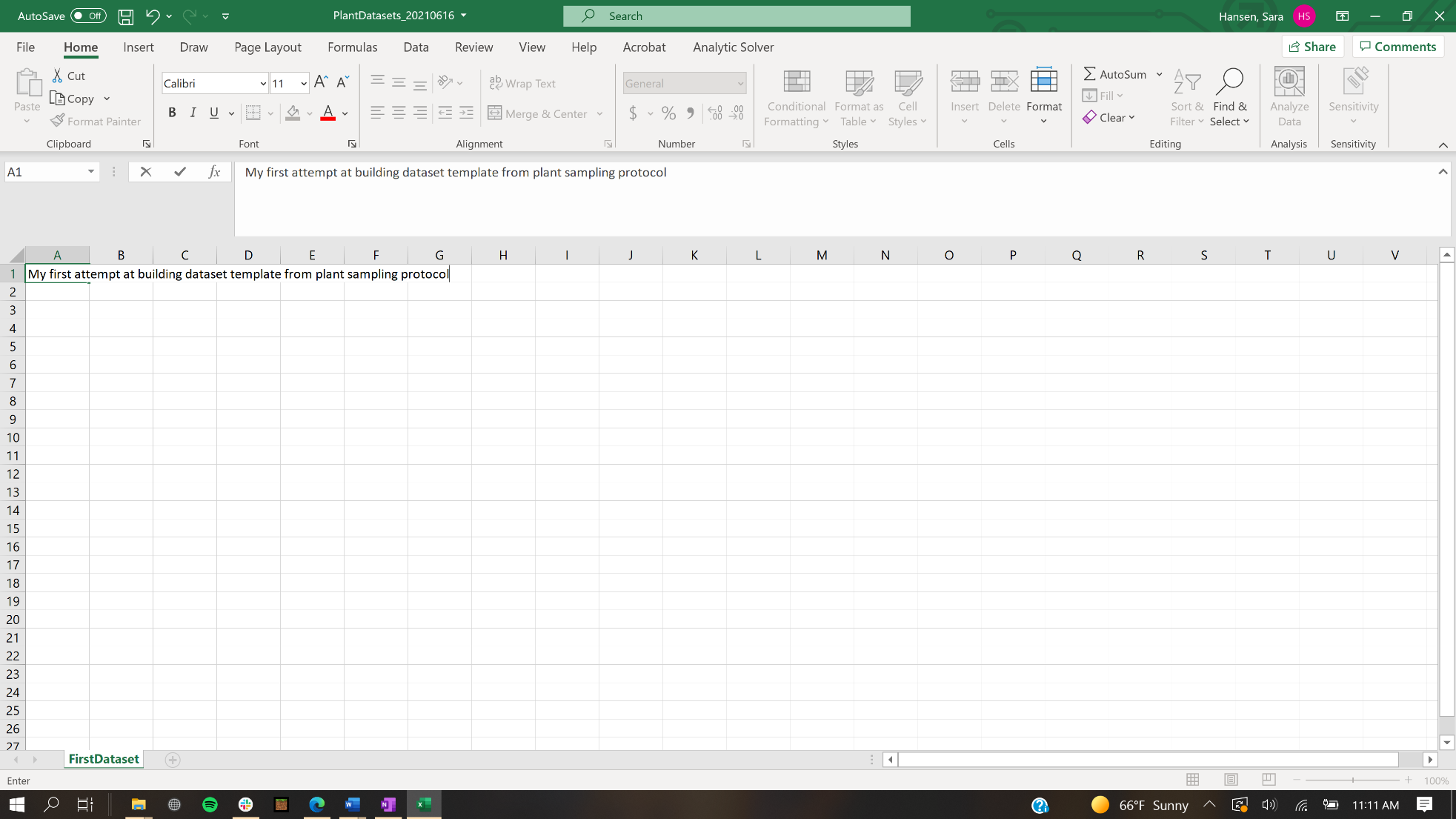
|  |
| --- |
| **List of fields (columns) to be collected**  *EFB flowering* |

1. Open a blank Excel workbook. Save it in your workspace (“BLUE\_BuildingDatasets\_yyyymmdd” folder) as “PlantDatasets\_yyyymmdd.”
2. Rename Sheet1 to “Dataset1” by right-clicking (Windows) or double-clicking (Mac) the sheet name, selecting Rename, and typing the new name.

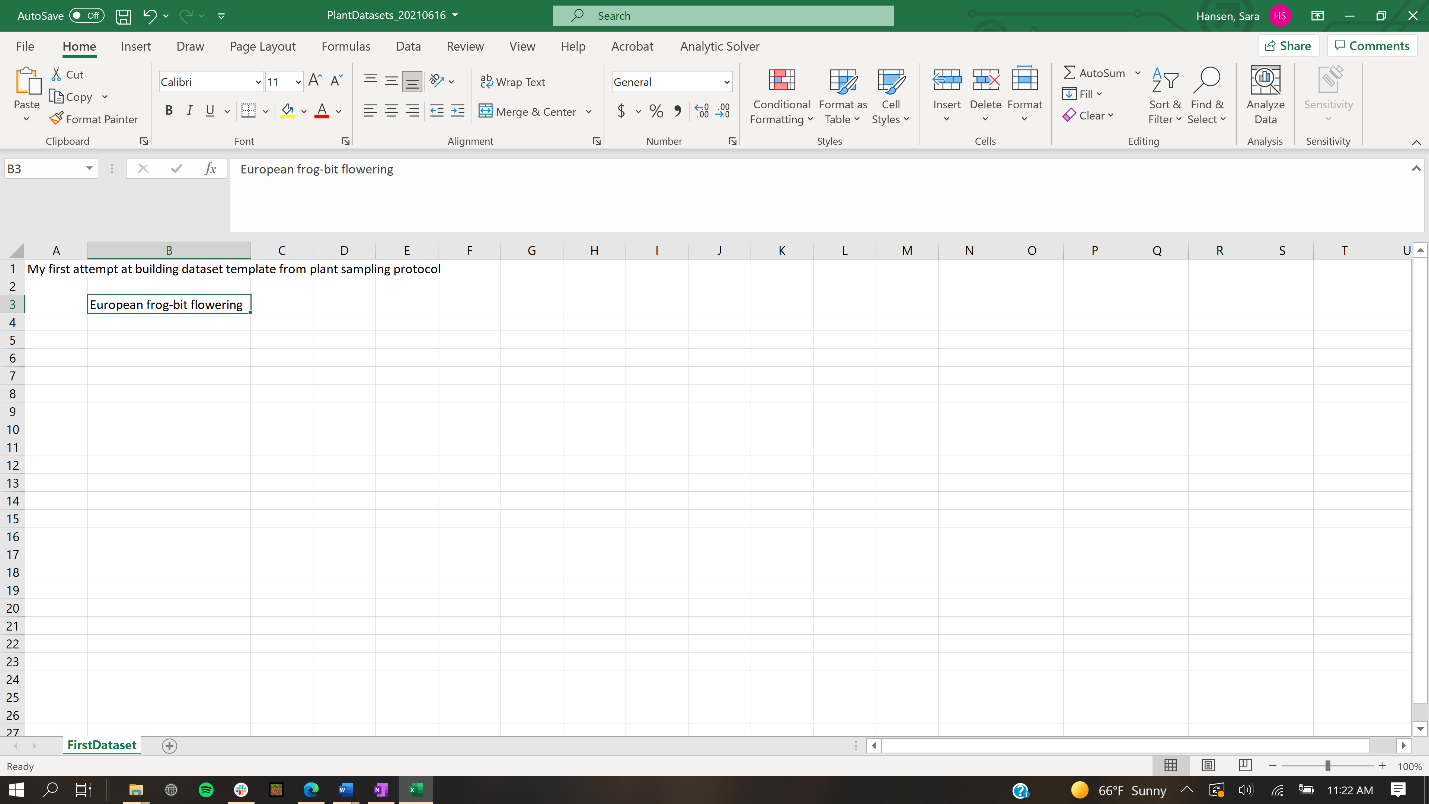
Graphical user interface, application, table

Description automatically generated

1. Type a short description of the sheet in cell A1.



1. Using your list of parameters to be collected from Step 1, type row and/or column names that represent the information (skipping one line after the description). Remember that columns should represent fields, or types of information, and rows should represent observations. In this case, consider each sampling plot an observation.



If your text runs over the cell, resize the column by double-clicking (Windows) or right-clicking (Mac) on the right side of the letter at the top. Create as many rows or columns as you think will be needed, and name them however you feel is appropriate. When you are finished with your template, don’t forget to save the Excel workbook.

**Activity III: Confirm That Your Template Includes All Necessary Fields**

Sometimes when you begin sampling, you realize your dataset template is not exactly how you need it for actual data collection. After beginning sampling, you may wish to change your dataset structure. That is perfectly okay! Just remember, you should do this at the very beginning, and once you settle on the structure you should keep it the same throughout the sampling. You should never have two separate datasets for the same protocol, so if you find it difficult to combine older and newer observations you might need to rethink the data structure to make sure all your observations match.

**Why build an Excel template before you begin sampling?**

It may be frustrating to find that the template you designed before sampling is not actually what you need for data collection in the field. However, it is very important to start with a template, even if you have to revisit and tweak it. Not only does it give you one place to store all your data, your template acts as a guide to make sure you collect every type of data you need to collect. There are several options other than Excel spreadsheets. For example, the ArcGIS Survey123 app (Esri 2021) allows you to list all the data types you need and set constraints on their values. In some cases, a printed data collection sheet is best, such as when you are working somewhere without an Internet connection or the ability to use an electronic device. No matter what method you use, your data will often end up in a tabular format, either by exporting them from a program like Survey123 or transcribing your written notes, so it is important to understand tabular structures and feel comfortable using them from start to finish.

Imagine you just sampled your first plot and want to make sure your template will work well for the rest of your plots. Because you haven’t yet finalized your template, you decide to record your data in your notebook first and plan to transcribe them immediately after finalizing your template. (*For the rest of your plots, you will have access to an electronic device and will record all your data directly in Excel.*)

*Procedure*

1. Review the notes from one sampling plot below.

On July 17, Alia visited Plot #90 located at 43.6822, -83.9159. The air temperature was 28 Celsius with 80% humidity. The water was 90 centimeters deep. About 11% of the plot area was covered in European frog-bit, which had only a few flowers. The duckweed species *Lemna minor* covered 5% of the plot. Starry stonewort (*Nitellopsis obtusa*) which is also an invasive species covered 30% of the plot and was flowering. No other species were observed.

Question: What new fields (columns) should you add to your dataset template that you didn’t include before?

1. Add a new sheet to your Excel workbook by clicking the “+” icon along the bottom. Name the new sheet “Dataset2.” Type a short description in cell A1, just as you did before.
2. You will now modify your original template based on the observations from one plot. This is a normal part of the process! Copy and paste your existing template in the “Dataset2” sheet. Add the new columns that you identified as missing based on the example sampling notes. (*Remember: Fields or data types should be stored in columns, and results or observations should be stored in rows. You should end up with one row of column names and one row of sampling results.*)

**Why did you create Dataset2 instead of just modifying Dataset 1?**

Keeping track of different versions of documents, data, and data templates is always good practice. By creating the new sheet Dataset2 and copying the contents of Dataset1, you are preserving Dataset1 as it was before and can now modify the template in Dataset2. Now you will be able to see what changes you had to make and can improve your template design skills for next time.

1. Fill in the example results from the notes above.

Question: You created the dataset template before knowing everything that would be collected. What problems did you face trying to translate the field notes to the template?

1. Your instructor will provide you with an example template and one row of example data. Review the example template with your instructor.

Question: How does this example template address some of the problems you identified above?

1. Create another Excel sheet in your workbook called “Dataset3,” and format it to exactly match the example. Notice all parameter names are the column headers, and their values are in the rows. Another important thing to notice is the **identifier** column, which in the example is “Plot number.” **Identifiers** link all information for a specific observation and allow you to easily reference the field values for that observation.

**Activity IV: Tidy Data**

You have created a clear, appropriate data collection template for your project. That is one very important step toward meaningful, refutable, and reproducible research.

It’s important to note there is no one right way to collect your data. There are many aspects to consider when designing and processing data, and many of them will change depending on methodology and project goals. However, a good rule of thumb is to make your data **tidy**. In a **tidy dataset** each column contains one variable, each row contains one observation, and it is clear what each cell value means. Because each column should contain only one variable, **tidy datasets are wider than they are long**, meaning they have more columns than rows.This rule may not always be true, but it is a good place to start.

Circle the tidy dataset(s) below, then list two reasons each one is or is not tidy.

**Dataset A**

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Profession and Annual Pay | Home city | Favorite things |
| Cassandra (surname unknown) | Electrician, $60,000 | Los Angeles, CA | Color = red, pet = dog |
| Williams, Thomas | Engineer, $86,000 | Orlando, FL | Cat, yellow, pizza |
| Laxmi Chopra (formerly Young) | Teacher, $44,000 | Detroit, MI | Favorite food is ice cream |
| Nam Ji-Won | Photographer, $95,000 | Chicago, IL | Loves pet fish |
| Diarmaid Argyll | Nurse, $72,000 | Austin, TX | Food = apples, color = blue |

Reasons:

**Dataset B**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Year | Month | Day | Item purchased | Quantity purchased | Per unit cost | Total cost |
| 2021 | 6 | 20 | Sweater | 1 | $19.98 | $19.88 |
| 2021 | 4 | 11 | Notebook | 5 | $2.50 | $12.50 |
| 2020 | 1 | 31 | Banana | 4 | $0.19 | $0.76 |
| 2019 | 8 | 22 | Computer | 1 | $699.00 | $699.00 |

Reasons:

**Dataset C** (*Hint: You are trying to figure out what each player scored in each round of the two games.*)

|  |  |  |  |
| --- | --- | --- | --- |
| Game | Player | Round | Score |
| Yahtzee | Alana | 1 | 220 |
| Yahtzee | E’Lise | 1 | 195 |
| Yahtzee | Jake | 1 | 155 |
| Yahtzee | Alana | 2 | 120 |
| Yahtzee | E’Lise | 2 | 175 |
| Yahtzee | Jake | 2 | 235 |
| Scrabble | Alana | 1 | 201 |
| Scrabble | E’Lise | 1 | 156 |
| Scrabble | Jake | 1 | 78 |
| Scrabble | Alana | 2 | 112 |
| Scrabble | E’Lise | 2 | 190 |
| Scrabble | Jake | 2 | 214 |

Reasons:

**Assessment**

1. You meet another scientist who believes they shouldn’t have to worry about organizing their data because they are more concerned with going out in the field and sampling organisms. How would you explain why planning and organizing for data are important to all scientists?
2. Fill in the blanks in the following sentences.
3. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ are the units of information.
4. In a dataset the fields, or types of information you collect, are stored in \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ and observations are stored in \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.
5. File names should avoid \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_, which are not machine-readable and will cause problems when programming.
6. A \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ dataset is usually wider than it is long, because many columns are needed to store individual types of information.
7. Now that you’ve translated field protocols to a tidy dataset template, you can create another template on your own. Using the turtle trapping protocol below, make a blank template using everything you’ve learned. (*Hint: Each trap is a separate observation.*)

**Protocol for Sampling Turtle Abundance in Saginaw Bay**

There are 10 segments (500-meters long) already marked.

3 traps will be deployed at each segment, for a total of 30 traps. Each trap will be set out for 5 days in August.

1. Locate the center of the 500-meter segment.

2. Deploy one trap at center and one at least 80 meters from the center on either side.

To deploy, tie one can of bait and flotation aids to metal trap, then anchor into substrate.

3. Measure water depth where trap is anchored.

4. Repeat steps 5-8 each day for the next 4 days.

5. Check every trap once per day and record the number of all organisms present.

6. Measure air temperature and humidity.

7. Note whether European frog-bit is present or absent in the 500-meter segment.

8. Reset trap and note the time.



Review the following checklist and make adjustments to your data template as needed.

File is stored in an appropriate location and named descriptively and consistently

Excel sheet includes brief description of template purpose and contents

All necessary information is represented by appropriate column names

Columns contain types of information (fields), so that each row will be one observation

Dataset is tidy and utilizes identifiers

**Glossary**

**Biodiversity**: Diversity of living things on Earth or in a particular area

*Humans threaten biodiversity by bringing about ecosystem changes that harm organisms and can cause their extinction.*

**Invasive species**: An introduced, nonnative species that is causing or likely will cause biological and/or economic harm

*Travel between continents may transport invasive species to new regions.*

**Data**: Units of information; provide context and record of knowledge

*Scientists communicate information using data, so it is important for all scientists to understand how they work.*

**Information**: Data in context of specific problem, question, or topic

*There is so much information available via the Internet, and much of it came from recorded data.*

**Knowledge**: Acquired understanding; often learned from information

*By completing this module, you will acquire knowledge of datasets and how to use them.*

**Fields**: In a dataset, the types of information collected; usually represented by columns

*Two fields that you should always include in your datasets are the date and collector.*

**Observations**: In a dataset, results of individual samples; usually stored in rows

*In the vegetation sampling protocol, the observations are the quadrats, so there are 96.*

**Dataset**: Collection of rows and columns with values representing combinations

*A dataset stores data in a way that is easily readable by humans and machines.*

**Identifier**: A tag attached to a specific record or observation that

*A person’s name is an identifier that is always linked to them and can be used to reference them.*

**Tidy data**: A data structure with each column representing one type of information, each row representing one observation, and cell values that can only have one meaning

*By making a dataset template before we start sampling, we can make sure we end up with tidy data that aren’t confusing to us later*

**Still have questions about creating tidy datasets? Here are some great resources to get you started:**

Hadley Wickham, Tidy Data: [Tidy Data (had.co.nz)](https://vita.had.co.nz/papers/tidy-data.pdf)

Towards Data Science, What is Tidy Data?: [What is Tidy Data?. A must-know concept for Data… | by Benedict Neo | Towards Data Science](https://towardsdatascience.com/what-is-tidy-data-d58bb9ad2458)

Data Carpentries, Data Organization: [Data Organization in Spreadsheets for Ecologists (datacarpentry.org)](https://datacarpentry.org/spreadsheet-ecology-lesson/)

U Chicago, Tidy Data: [Tidy data | Computing for the Social Sciences (uchicago.edu)](https://cfss.uchicago.edu/notes/tidy-data/)

**References:**

Brackett, M. 2013. *The Data – Information – Knowledge Cycle.* Dataversity. <https://www.dataversity.net/the-data-information-knowledge-cycle/>

Diagne, C., B. Leroy, A. C. Vaissière, R. E. Gozlan, D. Roiz, I. Jarić, J. M. Salles, C. J. A. Bradshaw, and F. Courchamp. 2021. High and rising economic costs of biological invasions worldwide. Nature 592: 571-576.

Dore, W. 1968. Progress of the European frogbit in Canada. Canadian Field-Naturalist 82: 76-84.

Esri, Inc. 2021. *ArcGIS Survey123*. Esri, Inc. <https://www.esri.com/en-us/arcgis/products/arcgis-survey123/overview>.

Minshall, W. H. 1940. Frog-bit *Hydrocharis morsus-ranae* L. at Ottawa. Canadian Field-Naturalist 54: 44-45.