

Name: _____

LA Initials:

Lab 5: The ggplot2 Visualization Challenge

Learning Objectives

By the end of this lab, students will be able to:

- Select and create appropriate ggplot2 graph types for common data structures (categorical, numerical, grouped, and time-based).
- Correctly map variables to ggplot2 aesthetics (x, y, fill, color, group) using unfamiliar datasets.
- Use external help resources (documentation, examples, online references) to independently solve plotting challenges.
- Diagnose and fix common ggplot2 errors through iterative refinement of code.
- Improve plot clarity and communication by applying at least one intentional visual refinement to each graph.

Getting started

Before you begin the lab activities, complete the following steps to make sure you are fully set up.

1. Get the Lab Worksheet.

Pick up a physical copy of the lab worksheet, or print one if you are working outside of class.
Download Lab Worksheet (PDF, if needed)

2. Open Posit Cloud and Start Lab 5.

Log in to **Posit Cloud** and navigate to the **course workspace**, and start the Lab 5 assignment.

3. Install required packages

Install the required packages for this lab, including **tidyverse**, **palmerpenguins**, and **gapminder** using the Packages pane or by running the following code *in the Console*:

```
install.packages(c("tidyverse", "lterdatasampler"))
```

4. Create an R Script

Create a new R script and save it as **lab-5-script.R**.

5. Load the packages

Copy and paste this code to your R script and run it to load the packages. Remember to run both library commands (Ctrl+Enter on one line, then the next).

```
# Load packages -----  
  
library(tidyverse)  
library(lterdatasampler)
```

i Checkpoint

At this point, you should have:

- These instructions open in a web browser.
- Your Lab 5 project open in Posit Cloud in another browser window.
- The required packages installed and loaded without errors
- The Lab 5 worksheet in front of you.

Do not continue until all of the above steps are working correctly.

Overview

In this lab, you will take on a series of **technical visualization challenges** using `ggplot2`. Unlike previous labs, you will be given **minimal instructions** and will be expected to rely on prior knowledge, careful reading of datasets, and effective use of help resources to complete each task.

You are **encouraged to work together** to troubleshoot problems, compare approaches, and learn from each other. Discussing errors, searching for solutions, and testing ideas collaboratively is an expected and valuable part of this lab.

Your goal is to create **one correct example of each of four major graph types** commonly used in biological data analysis: bar charts, histograms, boxplots, and time-series plots. Each challenge uses a different dataset and emphasizes choosing appropriate mappings and graph structures based on the data provided.

The **limited instructions are intentional**. This lab is designed to help you practice using help resources effectively, including classmates, LAs, course materials, and online documentation or examples. Optional extensions allow you to add visual polish if time permits.

Challenge 1: Bar Charts — Species Counts

Bar charts are used to compare **counts or totals across categories**. In ecology, they are commonly used to show the number of observations per species, taxonomic group, or site. The key idea is that the height (or length) of each bar represents how often something occurs.

For this challenge, you will use the `and_vertebrates` dataset from the `lterdatasampler` package. This dataset comes from the Andrews Forest Long-Term Ecological Research (LTER) site in Oregon and contains records of vertebrate observations, including amphibians, birds, mammals, and reptiles. Each row represents an observation of a species.

Your task is to create a **bar chart** showing the **count of observations for each species**. Your plot must meet the following requirements:

- Display species on the y-axis, with bar lengths representing counts.
- Add a title to the plot

If you finish early, consider refining the plot by reordering species by frequency, adding color, or adjusting labels for readability.

Challenge 2: Histograms — Air temperature distributions

Histograms are used to visualize the **distribution of a single numerical variable** by grouping values into bins and showing how frequently values occur within each range. They are especially useful for identifying skew, spread, and unusual values in environmental data.

For this challenge, you will use the `ntl_airtemp` dataset from the `lterdatasampler` package, which contains air temperature measurements collected at the North Temperate Lakes LTER site. Each row represents a single temperature observation recorded at a specific time.

Create a histogram showing the **distribution of air temperatures**. Your plot must meet the following requirements:

- Adjusted average air temperature must be mapped to the **x-axis**.
- The histogram must use an explicitly chosen **bin width** or number of **bins** (do not rely on the default).
- Both axis labels must be written in **sentence case** and include units where appropriate.

If you finish early, consider refining the plot by adjusting the bin width, adding a fill color with a subtle outline, or adding a vertical reference line (e.g., mean or median).

Challenge 3: Boxplots — Stem length across watersheds

Boxplots are used to compare the **distribution of a numerical variable across groups**. They summarize the median, spread, and potential outliers, making them especially useful for comparing variation among sites or treatments.

For this challenge, you will use the `hbr_maples` dataset from the `lterdatasampler` package, which contains measurements of sugar maple seedlings collected at the Hubbard Brook Experimental Forest. The dataset includes stem length measurements from seedlings sampled across multiple watersheds.

Create a boxplot showing **variation in stem length among watersheds**. Your plot must meet the following requirements:

- Watershed must be mapped to the **x-axis** and stem length to the **y-axis**.
- Axis labels must be clear, written in **sentence case**, and include units where appropriate.

If you finish early, consider refining the plot by adjusting box colors, overlaying individual data points, or reordering watersheds to improve readability.

Challenge 4: Time series with facets — Ice cover duration by lake

Time-series plots are used to show how a variable **changes over time**. When the same pattern needs to be compared across multiple groups, faceting allows each group to be shown in its own panel while keeping scales and formatting consistent.

For this challenge, you will use the `ntl_icecover` dataset from the `lterdatasampler` package. This dataset contains annual records of ice cover duration for multiple lakes at the North Temperate Lakes LTER site, including the year of observation and a lake identifier.

Create a time-series plot showing **ice cover duration over time**, split by lake using facets. Your plot must meet the following requirements:

- Year must be mapped to the **x-axis** and ice cover duration to the **y-axis**, using a **line plot**.
- Use faceting to create separate panels for each **lake (lakeid)**.
- Axis labels must be written in **sentence case** and include units where appropriate.

If you finish early, consider refining the plot by improving axis formatting, adjusting line thickness or color, or adding a smooth trend line to highlight long-term patterns within lakes.

Wrap-up and submission

This lab is graded for completion.

1. Make sure your script is saved in your project on Posit Cloud.
2. Show your graphs to an LA or instructor for a completion grade.

3. Keep your handout and **show it to a Learning Assistant for a completion grade before you leave lab.** You may do this as soon as you finish.