Module 2: Assignment 3

Ellen Bledsoe

2026-02-26

Assignment Description

Purpose

The goal of this assignment is to get started working on the final project for the course.

Task

Choose a dataset to work with for the final, write a research question, summarize and plot the data.

Criteria for Success

- Code is within the provided code chunks
- Code chunks run without errors
- Code produces the correct result
  - Code that produces the correct answer will receive full credit
  - Code attempts with logical direction will receive partial credit
- Written answers address the questions in sufficient detail

Due Date

March 5 at 12:15 pm

Assignment Questions

All questions are worth 1 point unless otherwise specified.

Explore Datasets

Click on the links for each dataset and read a little bit about each one.

- KrillBase
- Fur Seal Pups at Maiviken
- Gentoo Penguins on Bird Island
- Chinstrap Penguins on Signy Island

Now that you have an idea of what each data set is about, let's explore the actual data.

First, we need to load the `tidyverse` package.

```
# load the tidyverse
library(tidyverse)
```

Read all four datasets (.csv files) into RStudio using the `read_csv()` function (*not* `read.csv`). Remember to save each dataset as an object with a descriptive name.

```
krill <- read_csv("../../final_project/datasets/krill.csv")
fur_seals <- read_csv("../../final_project/datasets/fur_seals.csv")
gentoo <- read_csv("../../final_project/datasets/gentoos_and_temps.csv")
chinstrap <- read_csv("../../final_project/datasets/chinstraps_on_signy.csv")
```

1. Use a function (e.g., `head()` or `str()`) to look at the data in each data frame.

```
str(krill)
```

```
## spc_tbl_ [14,543 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Station          : chr [1:14543] "ake2008sars16" "ake2008sars17" "ake2008sars22" "ake2008sars23"
##  $ Date             : Date[1:14543], format: "2008-01-15" "2008-01-15" ...
##  $ Year             : num [1:14543] 2008 2008 2008 2008 2008 ...
##  $ Month            : num [1:14543] 1 1 1 1 1 2 2 2 2 2 ...
##  $ Day_Night        : chr [1:14543] "night" "day" "day" "day" ...
##  $ Krill_per_1m2    : num [1:14543] 0.116 0 1.454 38.784 305.411 ...
##  $ Krill_per_1m2_log: num [1:14543] 0.0477 0 0.3899 1.5997 2.4863 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Station = col_character(),
##   ..   Date = col_date(format = ""),
##   ..   Year = col_double(),
##   ..   Month = col_double(),
##   ..   Day_Night = col_character(),
##   ..   Krill_per_1m2 = col_double(),
##   ..   Krill_per_1m2_log = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(fur_seals)
```

```
## spc_tbl_ [3,601 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Date       : Date[1:3601], format: "2020-03-08" "2020-03-08" ...
##  $ Location   : chr [1:3601] "Tussock" "Tussock" "Tussock" "Tussock" ...
##  $ Sex        : chr [1:3601] "F" "F" "M" "M" ...
##  $ Weight_kg  : num [1:3601] 11.8 13.8 12.8 13.8 14.2 14.5 11.4 13.4 14 13.8 ...
##  $ Moult_score: num [1:3601] 3 3 3 3 3 3 3 3 3 3 ...
##  $ Year       : num [1:3601] 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 ...
##  $ Month      : chr [1:3601] "Mar" "Mar" "Mar" "Mar" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Date = col_date(format = ""),
##   ..   Location = col_character(),
##   ..   Sex = col_character(),
##   ..   Weight_kg = col_double(),
##   ..   Moult_score = col_double(),
##   ..   Year = col_double(),
##   ..   Month = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(gentoo)
```

```
## spc_tbl_ [173 x 5] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Species   : chr [1:173] "Gentoo penguin" "Gentoo penguin" "Gentoo penguin" "Gentoo penguin" ...
##  $ Year      : num [1:173] 1982 1982 1982 1982 1983 ...
```

```
##  $ Colony    : chr [1:173] "Johnson" "Mountain Cwm" "Natural Arch" "Square Pond" ...
##  $ Num_Chicks: num [1:173] NA NA NA NA 1880 767 829 282 636 NA ...
##  $ Dec_Temp_C: num [1:173] -13.1 -13.1 -13.1 -13.1 -15.7 -15.7 -15.7 -15.7 -12.8 -12.8 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Species = col_character(),
##   ..   Year = col_double(),
##   ..   Colony = col_character(),
##   ..   Num_Chicks = col_double(),
##   ..   Dec_Temp_C = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(chinstrap)
```

```
## spc_tbl_ [1,908 x 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ YEAR          : num [1:1908] 1996 1996 1996 1996 1996 ...
##  $ MONTH         : num [1:1908] 11 11 11 11 11 11 11 11 11 11 ...
##  $ DAY           : num [1:1908] 5 5 5 5 5 5 5 5 5 5 ...
##  $ BEAK_LENGTH_mm: num [1:1908] 45.2 47.8 49.1 48 48 43 44.7 53 49 45.4 ...
##  $ BEAK_DEPTH_mm : num [1:1908] 23.1 20.6 22.9 21.3 21.8 19.8 21.5 22.5 23.1 21.9 ...
##  $ SEX           : chr [1:1908] "M" "M" "M" "M" ...
##  $ MASS_kg       : num [1:1908] 4.5 4.3 5.4 5.4 4.7 4.7 4.6 5.7 4.8 4.9 ...
##  $ TEMP_C        : num [1:1908] 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.3 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   YEAR = col_double(),
##   ..   MONTH = col_double(),
##   ..   DAY = col_double(),
##   ..   BEAK_LENGTH_mm = col_double(),
##   ..   BEAK_DEPTH_mm = col_double(),
##   ..   SEX = col_character(),
##   ..   MASS_kg = col_double(),
##   ..   TEMP_C = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Choose Your Dataset

2. Now that you've explored the datsets, decide which one you would like to focus on for the final project. Below, tell us which one and why you find it interesting.

Answer:

Modifications

Find your dataset below.

I've made a few modifications to each of the dataset to make them a bit more manageable to work with. Based on how some of these variables will be used in future assignments, you should treat certain variables in the way I've listed below (even if they could be considered either *numeric* or *categorical).*

Krill

- You will want to use the `Krill_per_1m2_log` column as your dependent variable. This is a log-transformed version of the `Krill_per_1m2` column. You don't need to worry about *why* we transformed the data, but this column should be a bit easier to work with.

- The `Date`, `Year`, and `Month` columns should all be treated as *numeric* variables.
- While the `Station` column can *technically* be considered a categorical variable, please use the `Day_Night` column as your categorical variable for this assignment and for your research question.

Fur Seals

- If you choose to use the `Date` or `Year` column, treat them as *numeric* variables.
- On the other hand, if you use the `Month` column, treat it as a *categorical* variable.

Gentoo Penguins

- If you are using the `Year` column, treat it as a *numeric* variable.
- The dataset has a column for the previous year's December temperature. Because counts of chicks took place in January or February (and nests were counted in October/November), the number of chicks is most likely to be related to the temperatures directly after chicks hatched (December of the previous year).

Chinstrap Penguins

- I do not recommend using the `MONTH` or `DAY` column.
- If you use the `YEAR` column, treat it as a *numeric* variable.

Writing a Research Question

3. For each of the columns in your dataset, identify whether they contain categorical or numeric data.

   NOTE! Some data, such as dates (month, year), can act as both types of data. See the section above for clarification about columns in your chosen dataset.

   Type out the name of the column and then the descriptor. For example:

   - island_name: categorical
   - island_size_km: numeric

Answer: *Depends on the data set, see below:*

- *Krill: Station (categorical), Date (numeric), Year (numeric), Month (numeric), Day_Night (categorical), Krill_per_1m2 (numeric), Krill_per_1m2_log (numeric)*
- *Fur Seals: Date (numeric), Location (categorical), Sex (categorical), Weight_kg (numeric), Moult_score (either), Year (numeric), Month (categorical)*
- *Gentoo: Species (categorical), Year (numeric), Colony (categorical), Num_Chicks (numeric), Dec_Temp_C (numeric)*
- *Chinstrap: YEAR (numeric), MONTH (either), DAY (either), BEAK_LENGTH_mm (numeric), BEAK_DEPTH_mm (numeric), SEX (categorical), MASS_kg (numeric), TEMP_C (numeric)*

In order to write a successful research question, we need to identify our *dependent* variable and the *independent* variable that we think affects the dependent variable.

4. First, determine which variable is the *dependent* variable in your dataset. Remember, it must be *numeric*.

Answer:

- *Krill: Krill_per_1m2_log*
- *Fur Seals: Weight_kg, Moult_score*
- *Gentoo: Num_Chicks*
- *Chinstrap: BEAK_LENGTH_mm, BEAK_DEPTH_mm, MASS_kg*

5. Next, choose 1 *independent* variable from your dataset (each dataset has more than one from which to choose). This should be a variable that you believe might affect the values of the dependent variable. It can be either categorical OR numeric.

Answer:

- *Krill: Date, Year, Month, Day_Night (told above not to use Station)*
- *Fur Seals: Date, Location, Sex, Weight_kg (only if dependent is Moult_score), Year, Month (if dependent is Moult_score and they are treating month as numeric, give them feedback to treat month as categorical in the future)*
- *Gentoo: Year, Cology, Dec_Temp_C (Species is not a good option because it is only Gentoo)*
- *Chinstrap: YEAR, BEAK_LENGTH_mm, BEAK_DEPTH_mm, SEX, MASS_kg, TEMP_C (told above not to use MONTH or DAY)*

6. Now, let's put those 2 variables into a research question (see the last few slides of the "Aquaponics" lecture for a reminder of how to write a research question).

Answer: *Something along the lines of "Does [independent variable] impact [dependent variable]?"*

Summarize the Data

Let's get to know our data a little better.

7. Focus on your *dependent* variable. Calculate the minimum value (`min()`), maximum (`max()`) value, and average (`mean()`) value using the `tidyverse` functions (you might need `na.rm = TRUE`!).

- If your research question has a *categorical* independent variable, use the `group_by` function to find the above values of your dependent variable for each category of the independent variable
- If your research question has a *numeric* independent variable, calculate the minimum, maximum, and mean values for the independent variable in addition to the dependent variable

```
# NUMERIC independent variable
# dataframe %>%
#   summarize(min_dependent = min(dependent, na.rm = TRUE),
#             max_dependent = max(dependent, na.rm = TRUE),
#             avg_dependent = mean(dependent, na.rm = TRUE),
#             min_independent = min(independent, na.rm = TRUE),
#             max_independent = max(independent, na.rm = TRUE),
#             avg_independent = mean(independent, na.rm = TRUE))

## CATEGORICAL independent variable
# dataframe %>%
#   group_by(independent) %>%
#   summarize(min_dependent = min(dependent, na.rm = TRUE),
#             max_dependent = max(dependent, na.rm = TRUE),
#             avg_dependent = mean(dependent, na.rm = TRUE))
```

Plot the Data

Important Instructions and Info!

(a) You can use *any* variable from the dataset, not only the ones in your research question. In fact, you will *need* to use different variables to successfully create all of the plots.

(b) For *all* of the plots below, be sure to add:

- improved labels for the x-axis, y-axis, and the legend (if present)

- one of the following themes to your code: `theme_bw`, `theme_classic`, or `theme_light`.

(c) For each question, ***you get to choose the variables you use for each plot!*** Pay attention to which variables are numeric vs. categorical and where they should go in the code to produce the correct type of plot.

(d) Because I'm having you choose your own variables, there is no Answer Key for this assignment.

(e) Each plot is worth 2 points. Your interpretation is worth 1 point. Some of your plots may be hard to interpret. That's ok! In that case, explain why the plot is challenging to interpret.

STOP! Did you read the "Important Instructions and Info" section above?

8. Produce a histogram plot (just one, not multiple on the same plot).

   Write 1-2 sentences describing what information you can gather from plot. (2 points for plot, 1 for description)

```
# ggplot(data, aes(x = numeric_variable)) +
#   geom_histogram() +
#   labs(x = "Improved Label",
#        y = "Improved Label") +
#   theme_function()

# Krill: Krill_per_1m2_log, Date, Year, or Month
# Fur Seals: Date, Weight_kg, Moult_score, Year
# Gentoo: Year, Num_Chicks, Dec_Temp_C
# Chinstrap: YEAR, BEAK_LENGTH_mm, BEAK_DEPTH_mm, MASS_kg, TEMP_C
```

Answer:

9. Create a multiple histogram plot. Make sure we can see each histogram plot (make sure one isn't blocking another by modifying the transparency and position).

   Write 1-2 sentences describing what information you can gather from plot. (2 points for plot, 1 for description)

```
# ggplot(data, aes(x = numeric_variable, fill = categorical_variable)) +
#   geom_histogram(alpha = 0.5, position = "identity") +
#   labs(x = "Improved Label",
#        y = "Improved Label",
#        fill = "Improved Label") +
#   theme_function()

# Krill:
#   x-axis = Krill_per_1m2_log, Date, Year, or Month
#   fill/color = Day_Night
# Fur Seals:
#   x-axis = Date, Weight_kg, Moult_score, Year
#   fill/color = Location, Sex, Moult_score, Month
# Gentoo:
#   x-axis = Year, Num_Chicks, Dec_Temp_C
#   fill/color = Colony
# Chinstrap:
#   x-axis = YEAR, BEAK_LENGTH_mm, BEAK_DEPTH_mm, MASS_kg, TEMP_C
#   fill/color = SEX
```

Answer:

10. Create a box-and-whisker plot. Changing the color of the boxes is optional. Make sure to add points to the plot.

    Write 1-2 sentences describing what information you can gather from plot. (2 points for plot, 1 for description)

```
# NOTE: encourage them to put categorical variable on x,
# but do not take points off if it is on y)

# ggplot(data, aes(x = categorical, y = numeric, color = optional)) +
#   geom_boxplot() +
#   geom_jitter(alpha = 0.5, width = optional) +
#   labs(x = "Improved Label",
#        y = "Improved Label",
#        color/fill = optional) +
#   theme_function()

# Krill:
#   x-axis = Day_Night
#   y-axis = Krill_per_1m2_log, Date, Year, or Month
# Fur Seals:
#   x-axis = Location, Sex, Moult_score, Month
#   y-axis = Date, Weight_kg, Moult_score, Year
# Gentoo:
#   x-axis = Colony
#   y-axis = Year, Num_Chicks, Dec_Temp_C
# Chinstrap:
#   x-axis = SEX
#   y-axis = YEAR, BEAK_LENGTH_mm, BEAK_DEPTH_mm, MASS_kg, TEMP_C
```

Answer:

11. Create a scatter plot.

    Write 1-2 sentences describing what information you can gather from plot. (2 points for plot, 1 for description)

```
# NOTE: encourage them to put independent variable on x (if using),
# but do not take points off if it is on y)

# ggplot(data, aes(x = numeric_variable, y = numeric_variable) +
#   geom_point() +
#   labs(x = "Improved Label",
#        y = "Improved Label") +
#   theme_function()

# 2 of the following variables for each data frame
# Krill: Krill_per_1m2_log, Date, Year, or Month
# Fur Seals: Date, Weight_kg, Moult_score, Year
# Gentoo: Year, Num_Chicks, Dec_Temp_C
# Chinstrap: YEAR, BEAK_LENGTH_mm, BEAK_DEPTH_mm, MASS_kg, TEMP_C
```

Answer:

---

12. Do any of the plots you made above match up with your research question?

- If so, identify which plot and explain how you know it matches up correctly (based on numeric vs. categorical variables).
- If not, explain which plot type *would* match your research question and how you know (based on numeric vs. categorical variables).

Answer: This should describe the plot that uses the variables they chose for their research question; the plot should be based on whether they have chosen 1 numeric, 1 categorical or 2 numeric

Turning in Your Assignment

Follow these steps to successfully turn in your assignment on D2L.

1. Click the `Knit` button up near the top of this document. This should produce a PDF file that shows up in the `Files` panel on the bottom-right of your screen.
2. Click the empty box to the left of the PDF file.
3. Click on the blue gear near the top of the `Files` panel and choose Export.
4. Put your last name at the front of the file name when prompted, then click the Download button. The PDF file of your assignment is now in your "Downloads" folder on your device.
5. Head over to D2L and navigate to appropriate assignment. Submit the PDF file that you just downloaded.