Module 2, Assignment 2

Ellen Bledsoe

2023-02-23

Assignment Details

Purpose

The goal of this assignment is to practice developing successful workflows to get from the starting data to a plot.

Task

Write R code which produces the correct answers and reflect on workflow.

Criteria for Success

- Code is within the provided code chunks
- Code is commented with brief descriptions of what the code does
- 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 2 at midnight MST

Assignment Questions

Our goal for this assignment is to start with a data frame of data, summarize the data in constructive ways, and plot the data to answer some questions. To get from Point A to Point B to Point C requires some planning.

In this assignment, we will make a plan, execute how we would actually go about the process, and then evaluate how well our original plan matches with the path we actually used.

Data Summary

The aquaculture scientists on Team Antarctica have been working on developing a new diet for tilapia based on soy-protein, and they are interested in whether incorporating this into fish diets will result in faster growth rates.

They also want to know if those growth rates are related to average tank temperature.

They've provided us with the data below to analyze.

First, let's take a look at the data we will be using for the assignment.

1. As always, we start by loading the `tidyverse`. (1 point)

```
library(tidyverse)
```

2. Next, we need to load in our tilapia growth dataset. Call the data frame `growth`. Then use both the `head` and `tail` functions to take a look at the data. (1 point)

```
growth <- read_csv("../data/tilapia_growth.csv")
```

```
## Rows: 320 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): tank_category
## dbl (5): tank_id, fish_id, perc_soy_protein, day_30_weight, avg_tank_temp
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(growth)
```

```
## # A tibble: 6 x 6
##   tank_id fish_id perc_soy_protein day_30_weight avg_tank_temp tank_category
##     <dbl>   <dbl>            <dbl>         <dbl>         <dbl> <chr>
## 1       1       1              0.2          334.          77.2 warm
## 2       1       2              0.2          198.          77.2 warm
## 3       1       3              0.2          315.          77.2 warm
## 4       1       4              0.2          316.          77.2 warm
## 5       1       5              0.2           89.4         77.2 warm
## 6       1       6              0.2           74.7         77.2 warm
```

```
tail(growth)
```

```
## # A tibble: 6 x 6
##   tank_id fish_id perc_soy_protein day_30_weight avg_tank_temp tank_category
##     <dbl>   <dbl>            <dbl>         <dbl>         <dbl> <chr>
## 1      16     315              0.8          1228.          76.1 warm
## 2      16     316              0.8           630.          76.1 warm
## 3      16     317              0.8           508.          76.1 warm
## 4      16     318              0.8           443.          76.1 warm
## 5      16     319              0.8           495.          76.1 warm
## 6      16     320              0.8          1078.          76.1 warm
```

3. Before we can plan any strategy, we need to understand the data that we have. Answer the following questions about the data. (0.5 points each, 2 points total)

   a. What does one row represent? (One tank? One temperature? One treatment? One fish?)

   b. How many tanks of fish were sampled?

   c. How many fish were sampled?

   d. How many different "percent soy protein" levels (treatments) are there? In this assignment, we will treat `perc_soy_protein` as a *categorical* variable, so another way to ask this question is: how many categories of percent soy protein do we have?

Our Task

We have been asked by our aquaculture specialists to provide them with the following:

a. a data frame with the average growth (`day_30_weight`) per treatment (`perc_soy_protein`)
b. a boxplot plot that shows the relationship (or lack thereof) between percent soy protein in the diet and the weight at 30 days
c. a data frame with the average growth per treatment (`perc_soy_protein`) *and* if tanks are warm or cold
d. a multiple scatter plot that shows the relationship (or lack thereof) between average tank temperature, the weight at 30 days, and the percent soy in the diet.

They've also asked that we provide all weights in kilograms instead of grams (the `day_30_weight` is currently in grams).

Prediction

4. Spend some time thinking about each one of these steps. What steps will you need to take to produce the end result? What data frame will you use? What columns will you use? What functions will you use? How will you plot things?

   For each of the 4 tasks listed above, *describe* (do NOT code) how you will get from the starting point (a data frame) to the result (another data frame or a plot).

   This question will be grade ***only on completion***, not on whether or not your plan is *correct*. (2 points)

   Task (a):

   Task (b):

   Task (c):

   Task (d):

Execution

Now let's actually go ahead and complete our tasks with code.

FIRST! We need to run the line of code below. It's a little wonky, but for the rest of the assignment to work the way I want it to, we need to tell R that we want to treat the `perc_soy_protein` as a categorical variable.

```
growth <- growth %>%
  mutate(perc_soy_protein = as.factor(perc_soy_protein))
```

Because the aquaculture team has asked for everything to use kilograms instead of grams, it makes sense for us to add a `day_30_weight_kg` column before we tackle any of the specific tasks.

5. Use the `mutate()` function to add a `day_30_weight_kg` column to the growth data frame. (Hint: divide grams by 1000 to get kg)

   Remember to "overwrite" the growth data frame (use the assignment operator) so the new column is permanently added to the `growth` data frame for us to use in the rest of the assignment. (2 points)

```
growth <- growth %>%
  mutate(day_30_weight_kg = day_30_weight / 1000)
```

Task (a)

A data frame with the average growth (`day_30_weight_kg`) per treatment (`perc_soy_protein`)

6. Create a new data frame called `growth_by_treatment`. We first want to group the data by treatment (`perc_soy_protein`) then calculate the average weight in kg. (2 points)

```
growth_by_treatment <- growth %>%
  group_by(perc_soy_protein) %>%
  summarize(mean_weight_kg = mean(day_30_weight_kg))
growth_by_treatment
```

```
## # A tibble: 4 x 2
##   perc_soy_protein mean_weight_kg
##   <fct>                     <dbl>
## 1 0.2                       0.276
## 2 0.4                       0.441
## 3 0.6                       0.618
## 4 0.8                       0.829
```

Task (b)

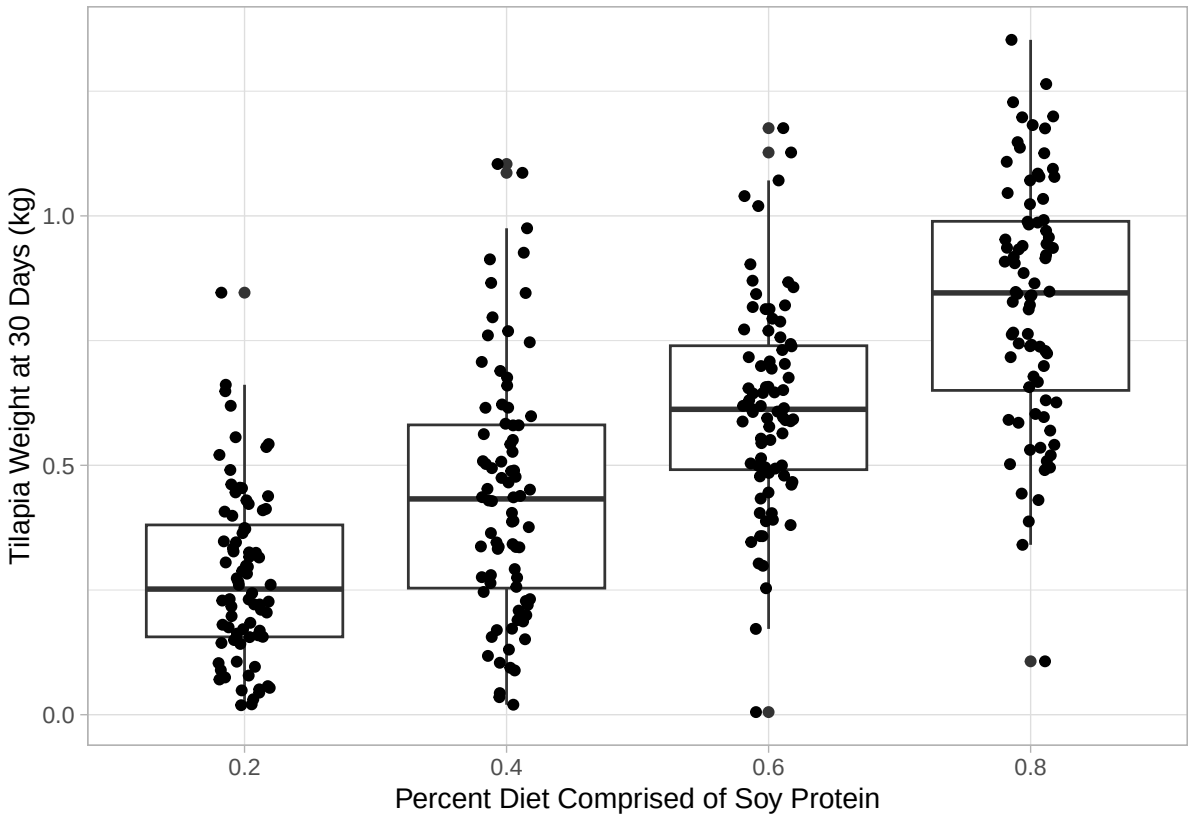
A box plot that shows the relationship (or lack thereof) between percent soy protein in the diet and the weight at 30 days

Based on the values we calculated in task (a), it looks like there is probably a positive relationship between the percent of soy protein in tilapia diet and growth. Let's plot the data to confirm.

7. Make a box plot with `perc_soy_protein` on the x-axis (horizontal) and `day_30_weight_kg` on the y-axis (vertical). Change the axis labels to be more easily understood and add a theme. (2 point)

   (Hint: because we want to plot *all* of the values, not just the mean values, we need to use the original `growth` data frame)

```
# a few people might have a scatter plot because I made a mistake with the wording; that's fine
ggplot(growth, aes(perc_soy_protein, day_30_weight_kg)) +
  geom_boxplot() +
  geom_jitter(width = 0.1) +
  labs(x = "Percent Diet Comprised of Soy Protein",
   y = "Tilapia Weight at 30 Days (kg)") +
  theme_light()
```



Task (c)

A data frame with the average growth per treatment (`perc_soy_protein`) *and* if tanks are warm or cold

8. Now that we have our `tank_category` column, we can use it in our `group_by` function, along with `perc_soy_protein`, to calculate the average `day_30_weight_kg` for warm and cold tanks in each treatment. Call this new data frame `growth_by_temp_treatment`. (2 points)

```
growth_by_treatment_and_temp <- growth %>%
  group_by(perc_soy_protein, tank_category) %>%
  summarise(mean_weight_kg = mean(day_30_weight_kg))
```

```
## `summarise()` has grouped output by 'perc_soy_protein'. You can override using
## the `.groups` argument.
```

```
growth_by_treatment_and_temp
```

```
## # A tibble: 8 x 3
## # Groups:   perc_soy_protein [4]
##   perc_soy_protein tank_category mean_weight_kg
##   <fct>            <chr>                 <dbl>
## 1 0.2              cold                  0.279
## 2 0.2              warm                  0.272
## 3 0.4              cold                  0.420
## 4 0.4              warm                  0.448
```

```
## 5 0.6              cold                0.617
## 6 0.6              warm                0.620
## 7 0.8              cold                0.824
## 8 0.8              warm                0.833
```

Task (d)

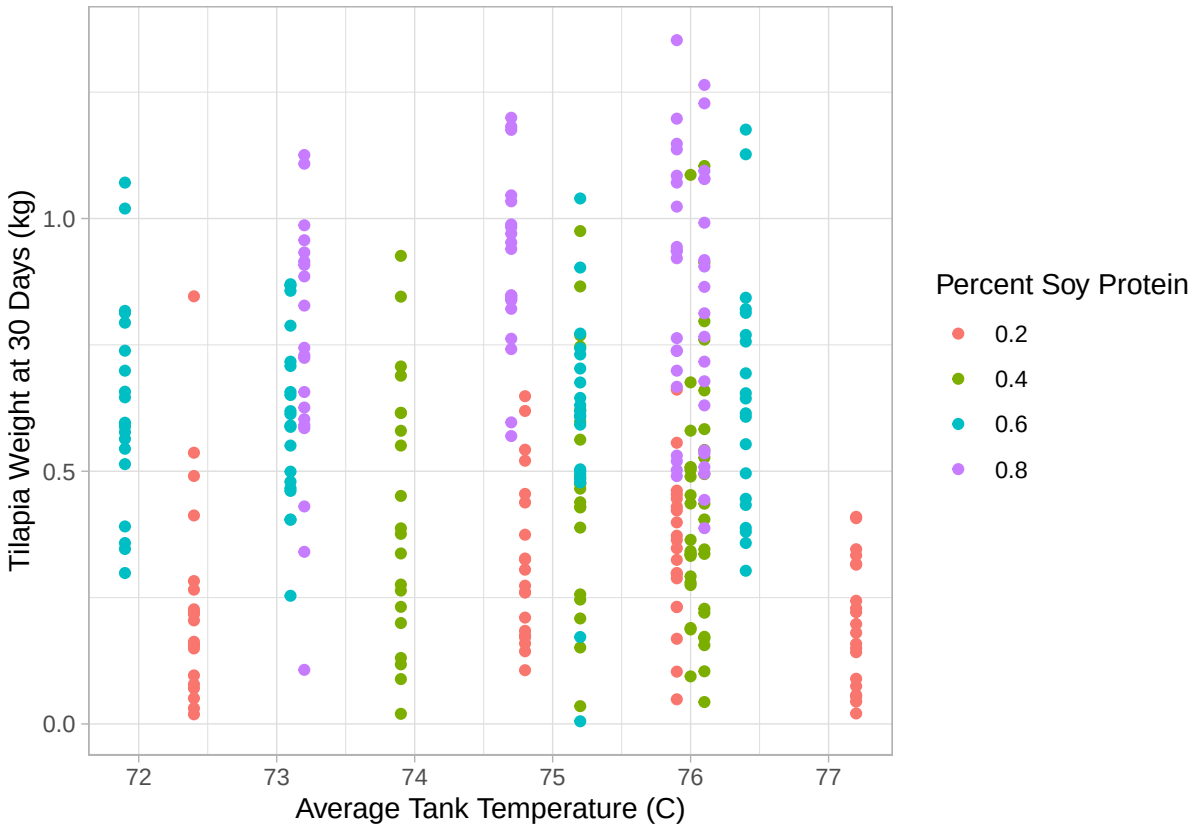
A scatter plot that show the relationship (or lack thereof) between average tank temperature and the weight at 30 days

Based on our results above, do we think the tank being warm or cold has much of an influence? Let's plot our data to confirm.

9. Make a multiple scatter plot with `avg_tank_temp` (not `tank_category`!) on the x-axis (horizontal) and `day_30_weight_kg` on the y-axis (vertical). Change the color so it represents the percent soy protein. Change the axis labels to be more easily understood and add a theme. (2 point)

   (Hint: because we want to plot *all* of the values, not just the mean values, we need to use the original `growth` data frame)

```
ggplot(growth, aes(avg_tank_temp, day_30_weight_kg, color = perc_soy_protein)) +
  geom_point() +
  labs(x = "Average Tank Temperature (C)",
       y = "Tilapia Weight at 30 Days (kg)",
       color = "Percent Soy Protein") +
  theme_light()
```



As suspected, there doesn't seem to be much of a difference based on tank temperatures.

Reflection

10. Imagine we had decided to treat `perc_soy_protein` as a continuous variable. What type of plot we have used for task (b)? Would we have been able to complete task (d)? Why or why not? (2 points)

    Answer: Yes or no is acceptable. I have not taught them how to plot 3 continuous variables on a scatter plot, but it is doable. If they say no, you can only plot 2 continuous and 1 categorical, that is valid. If they say you *can* plot 3 because you can use color as a spectrum, that is also viable. Just make sure they explain why it would (or would not) work.

11. Write 3-5 sentences about if and how your predictions and execution differed and what you learned through the process. (3 points)

    Examples of questions to answer: How did your initial prediction of how you expected to accomplish the 4 tasks match up with how we actually went about doing it? Were they similar? Were there common mistakes that you made beforehand? Did you plan a different execution from what we did above that you think would also work?

    Answer: