Module 2, Assignment 1

Ellen Bledsoe

2023-09-24

Assignment Details

Purpose

The goal of this assignment is to assess your ability to produce and interpret scatter plots in both base R and `ggplot2`.

Task

Write R code which produces the correct answers and text to correctly interpret the plots produced.

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

Oct 3 at 4pm MST

Assignment Questions

For this assignment, we are going to be making plots! We are going to use a data set containing data from a sample of our fish tanks. The data contains information about the tanks sampled and how many sick fish are contained in each tank.

First, let's prepare our data.

1. Load the `tidyverse` package into the work space.

```
library(tidyverse)
```

2. Read in the data using the `read_csv()` function. Name the data frame `sick_fish`.

```
sick_fish <- read_csv("../data/fish_sick_data.csv")
```

```
## Rows: 50 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): species
## dbl (6): tank_id, avg_daily_temp, num_fish, day_length, tank_volume, num_sick
##
## 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.
```

3. Use the `head()` function to take a look at the columns in the data frame.

```
head(sick_fish)
```

```
## # A tibble: 6 x 7
##   tank_id species avg_daily_temp num_fish day_length tank_volume num_sick
##     <dbl> <chr>            <dbl>    <dbl>      <dbl>       <dbl>    <dbl>
## 1     388 tilapia           24.3       93         10        399.        3
## 2     425 tilapia           24.6       98         11        400.        4
## 3     420 tilapia           23.0      103          9        399.        2
## 4     819 trout             14.1       85         11        401.       14
## 5     176 tilapia           23.3       98         10        400.        3
## 6     926 trout             13.8       79         12        400.       10
```

4. How many rows does the data frame have? How many columns?

   Answer: 50 rows, 7 columns

5. Take a look at the data frame.

   - What does one row (observation) represent (e.g., an individual fish?, all fish of a certain species? all tanks of a certain species? etc.)? (1 point)

   Answer: one row represents one fish tank that was sampled

   - Based on the column names, briefly (just a few words) describe what data is included in each column. (2 points)

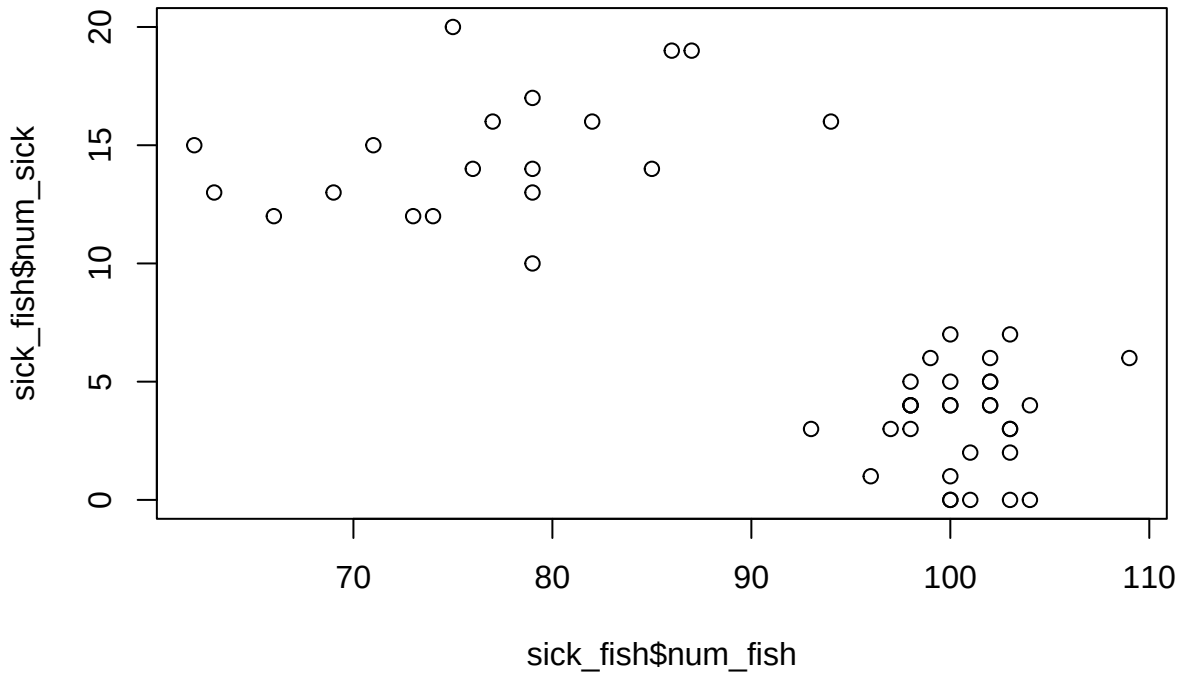
   Answer: tank_id = identifier for individual tanks; species = fish species in the tank; avg_daily_temperature = average water temp in the tank; num_fish = total number of fish in the tank; day_length = amount of daylight provided for tanks; tank_volume = water volume in the tank; num_sick = number of sick fish per tank

6. Take a look at the columns that have the total number of fish in the tank and the number of sick fish per tank. Determine whether these two columns are continuous or categorical.

   Answer: both num_fish and num_sick are continuous

7. Using base R, create a scatter plot. Put the total number of fish on the x-axis and the number of sick fish on the y-axis.

```
plot(sick_fish$num_fish, sick_fish$num_sick)
```



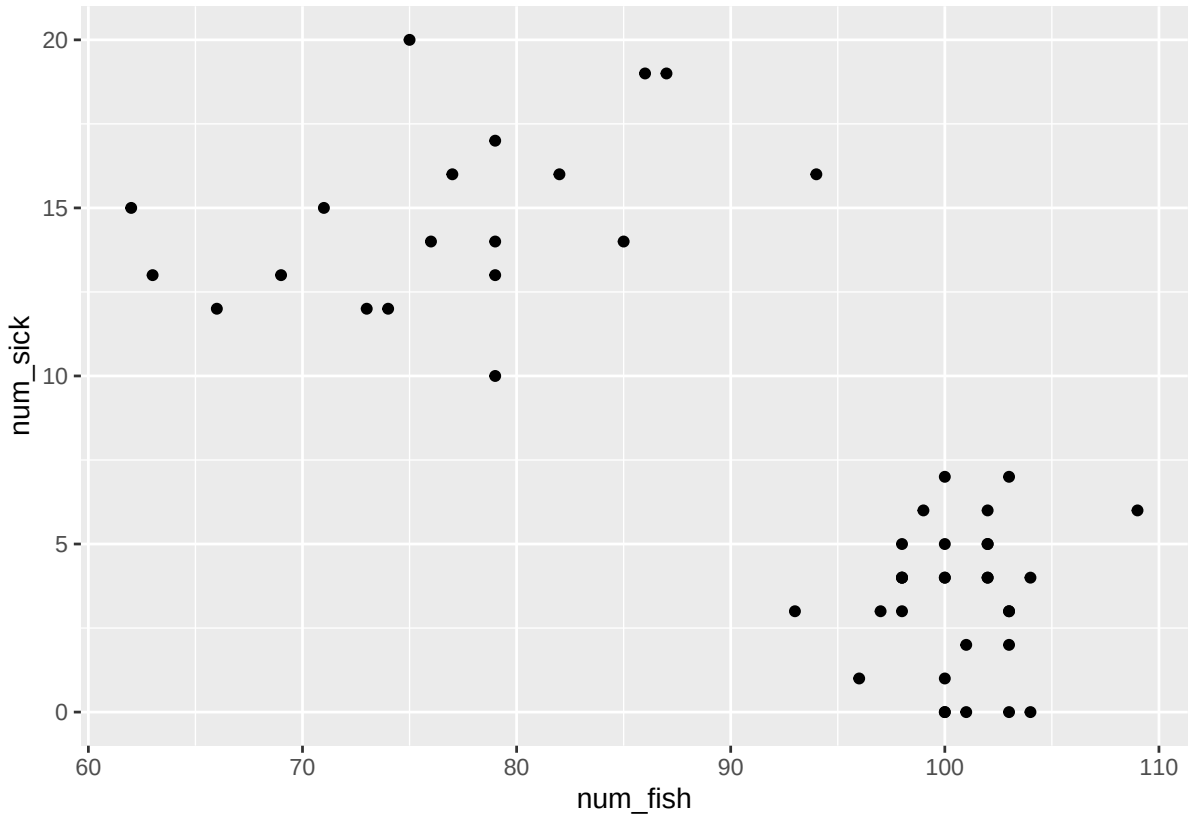
8. Interpret the scatter plot. What is the plot telling you? Is there a positive or negative relationship between the two variables. What does that mean? (2 points)

   Answer: negative relationship, meaning that the more fish that are in the tank, the fewer sick fish there are (which is counter-intuitive).

Now, let's use what we've learned about `ggplot2` to recreate the same scatter plot. We will do this in 2 steps. In the first step, we will add the data, axes, and geom. In the second step, we will modify the plot to increase interpretability by renaming axes and adding a theme.

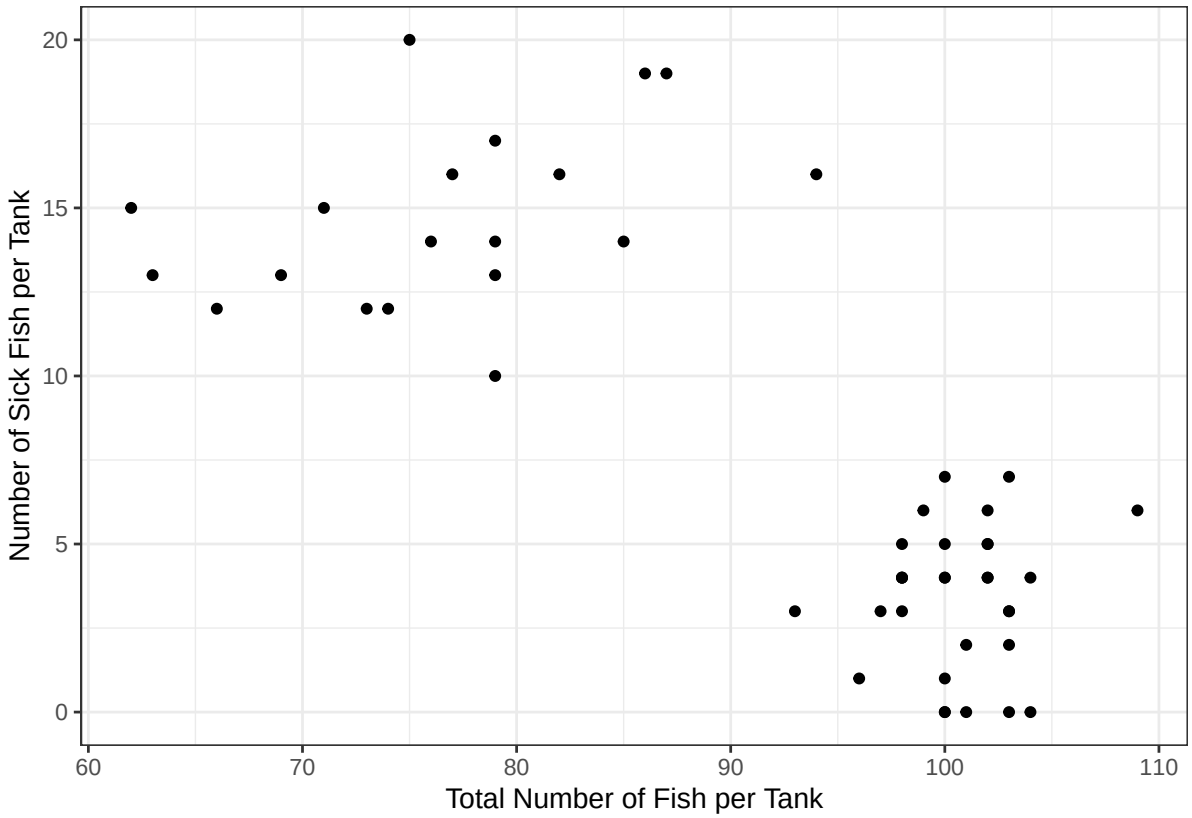
9. Create the scatter plot using the proper `geom` function.

```
ggplot(sick_fish, aes(num_fish, num_sick)) +
  geom_point()
```



10. Now, add to the plot that we created above to make it clearer to understand. Add better axes labels and choose a theme. (2 point)

```
ggplot(sick_fish, aes(num_fish, num_sick)) +
  geom_point() +
  labs(x = "Total Number of Fish per Tank",
       y = "Number of Sick Fish per Tank") +
  theme_bw()
```



Looking at the scatter plots, there seem to be two distinct groups. Let's investigate this a bit further.

11. For *each* fish species, calculate the average and the standard deviation for the number of sick fish. (2 points)

    Hint: think back to Module 1 and which function we can use to "split, apply, combine"

```
sick_fish %>%
  group_by(species) %>%
  summarise(mean_sick = mean(num_sick),
        sd_sick = sd(num_sick))
```

```
## # A tibble: 2 x 3
##   species mean_sick sd_sick
##   <chr>       <dbl>   <dbl>
## 1 tilapia      3.39    2.11
## 2 trout       14.7     2.68
```

12. Based on the summary data you've calculated above, do you think the two clumps of data correspond to the two species? Which species seems to be the one that is causing the most problems? (2 points)

    Answer: seems like trout are driving the issues; seems reasonable that clumps could belong to species

13. Do some brainstorming. Based on the plots we've made and the calculations we've done so far, are we *sure* that the two clumps of data we see in the data visualization do, in fact, correlate to the two different species? How could we confirm? How might we include that information in the plot? (2 points)

    Answer: many answer options here. could take average of total num fish to see if that corresponds to what is on the plot; could change some characteristic of the plot (point shape, color, size) to represent fish species; could add labels (would get busy)