Module 2 Wrap-Up

Ellen Bledsoe

2023-02-28

What is causing the food poisoning?

At the beginning of Module 2, we set out to discover what was causing food poisoning among our colleagues at our Antarctic base. Let's put everything we've learned about descriptive statistics and data visualization to use to try to hunt down what the problem is.

Set-up

Let's load the package and data we will need.

Packages

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Data

We've already explored the tank data provided to us from our aquatic specialists. Unfortunately, those data didn't really give us what we needed to answer our question. What we really need is data on where the sick fish are coming from! From there, we can hopefully narrow down what information about those tanks different from other tanks.

It would be impossible to sample all 1000 tanks for sick fish on our timeline, so we asked our aquatic specialists to randomly choose 50 tanks to sample for sick fish. This is the data in the `fish_sick_data.csv` file.

```
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.
```

Let’s check out our data and see what we are working with.

```
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
```

Which Fish?

A good first step would be to narrow down whether tilapia or trout are the likely culprits.

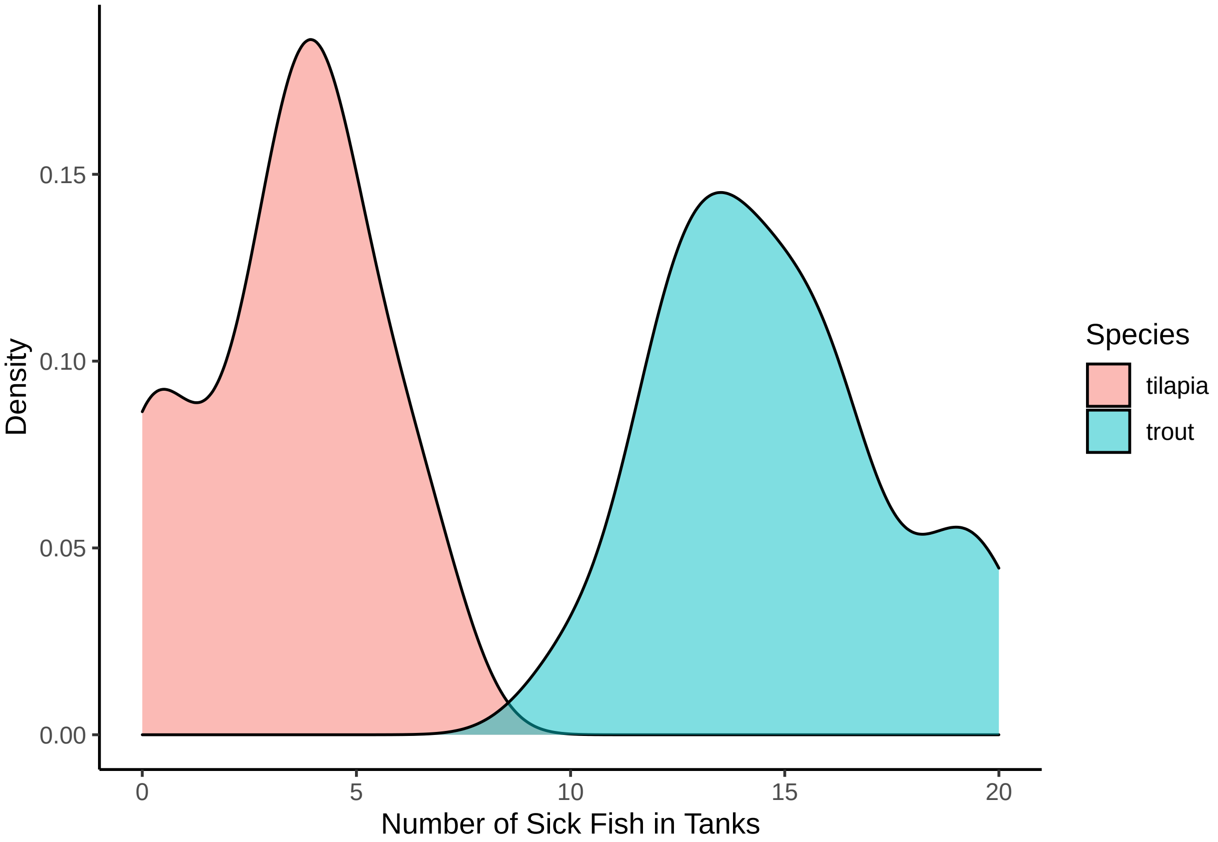
To start, let’s calculate the mean and standard deviation of sick fish per species.

```
sick_fish %>%
  group_by(species) %>%
  summarize(mean_sick_fish = mean(num_sick),
            sd_sick_fish = sd(num_sick))
```

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

Wow, trout aren’t looking so good! Let’s plot the data to see what’s happening. We want a plot that will allow us to compare the data between the two groups of data (fish species). We actually have a few options! Let’s go with a multiple density plot this time.

```
ggplot(sick_fish, aes(num_sick, fill = species)) +
  geom_density(alpha = 0.5) +
  labs(x = "Number of Sick Fish in Tanks",
       y = "Density",
       fill = "Species") +
  theme_classic()
```



Density

Wait a second! Take a look back at the data. There is a "number of fish" column, indicating the total number of fish in the tank, and it looks like those numbers can differ pretty widely. We should probably take into account how many fish there are in the tank to begin with. 12 sick fish out of 50 is probably a bigger deal than 12 sick fish out of 100! What we need to do is calculate a density of fish–number of sick fish / number of total fish.

```
# take into account the number of fish in the tank: density of sick fish
sick_fish <- sick_fish %>%
  mutate(density = num_sick/num_fish)
```

Small Groups Let's make sure our conclusions about trout being the true culprits still hold when we account for the total number of fish in the tank.

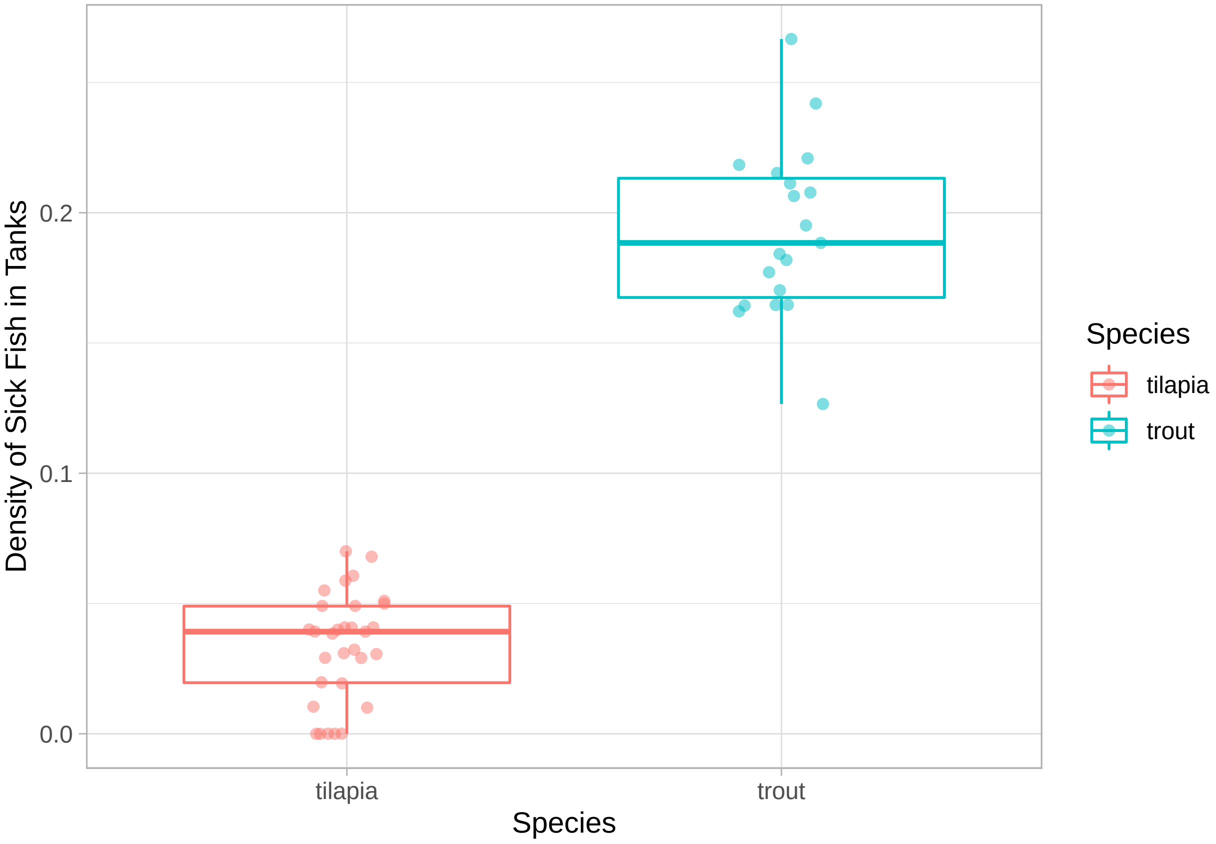
Work in small groups to do the following:

- find the average number and standard deviation of sick fish for both species
- make a plot that compares the distributions of sick fish numbers for both species (you have multiple options here!)

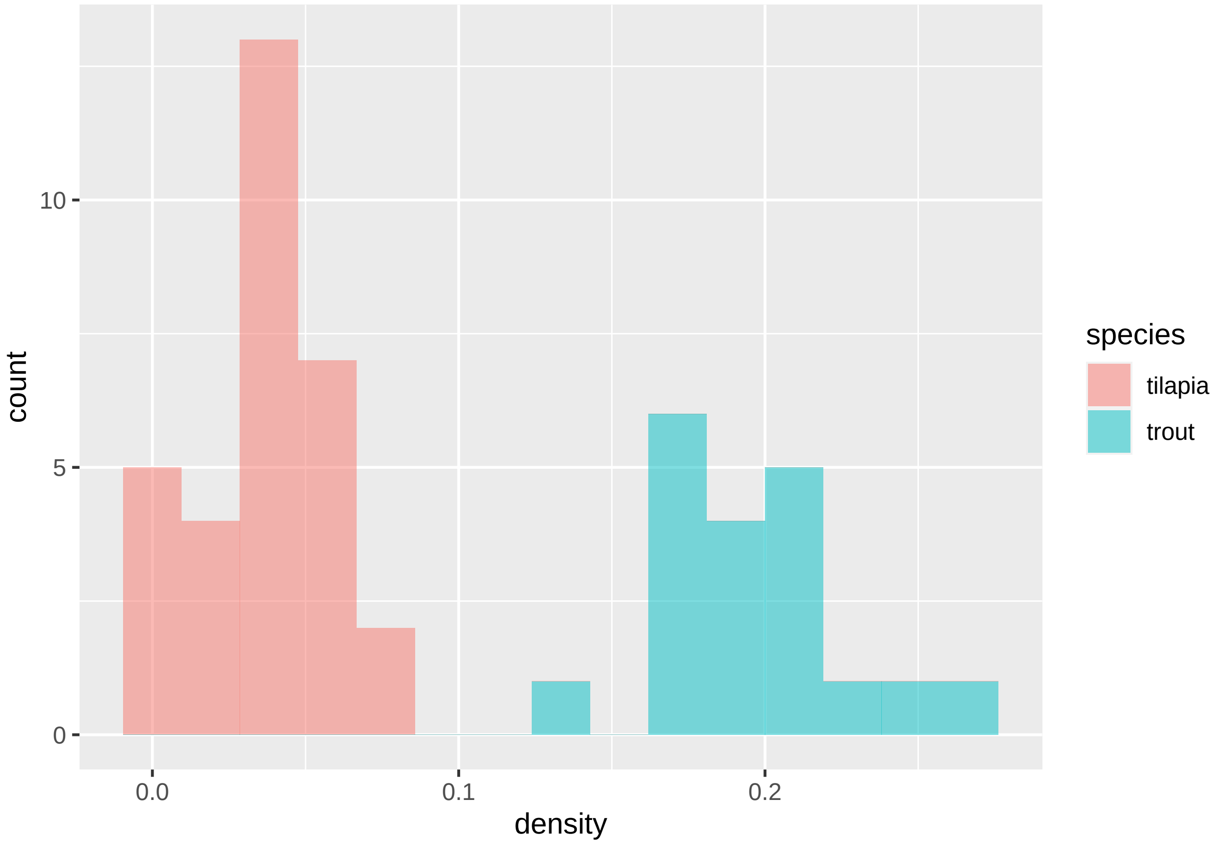
```
sick_fish %>%
  group_by(species) %>%
  summarize(mean_sick_fish = mean(density),
            sd_sick_fish = sd(density))
```

```
## # A tibble: 2 x 3
##   species mean_sick_fish sd_sick_fish
##   <chr>            <dbl>        <dbl>
## 1 tilapia         0.0336       0.0207
## 2 trout           0.193        0.0327
```

```
ggplot(sick_fish, aes(species, density, color = species)) +
  geom_boxplot() +
  geom_jitter(alpha = 0.5, width = 0.1) +
  labs(x = "Species",
       y = "Density of Sick Fish in Tanks",
       color = "Species") +
  theme_light()
```



```
# histogram
ggplot(sick_fish, aes(density, fill = species)) +
  geom_histogram(alpha = 0.5, bins = 15)
```



Uh oh…the trout densities look even worse than just the number of sick fish. We need to take a closer look at what is going on in the trout tanks!

We should create a data frame that only contains trout to work with for the rest of our analyses. Take a few minutes to work on that; call it `sick_trout`.

```
sick_trout <- sick_fish %>%
  filter(species == "trout")
```

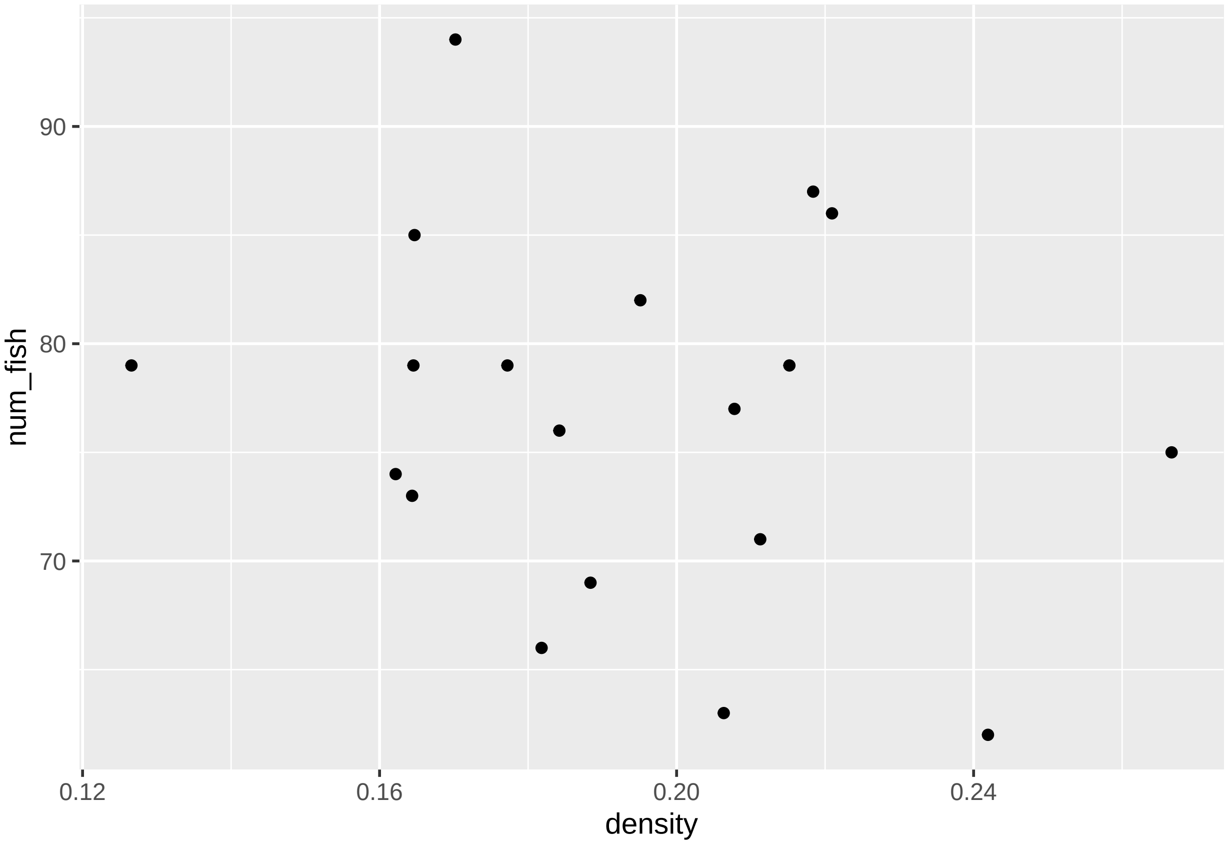
What Environmental Factor?

Take a look back at the data frame. Which columns are environmental variables that could be driving the issues?

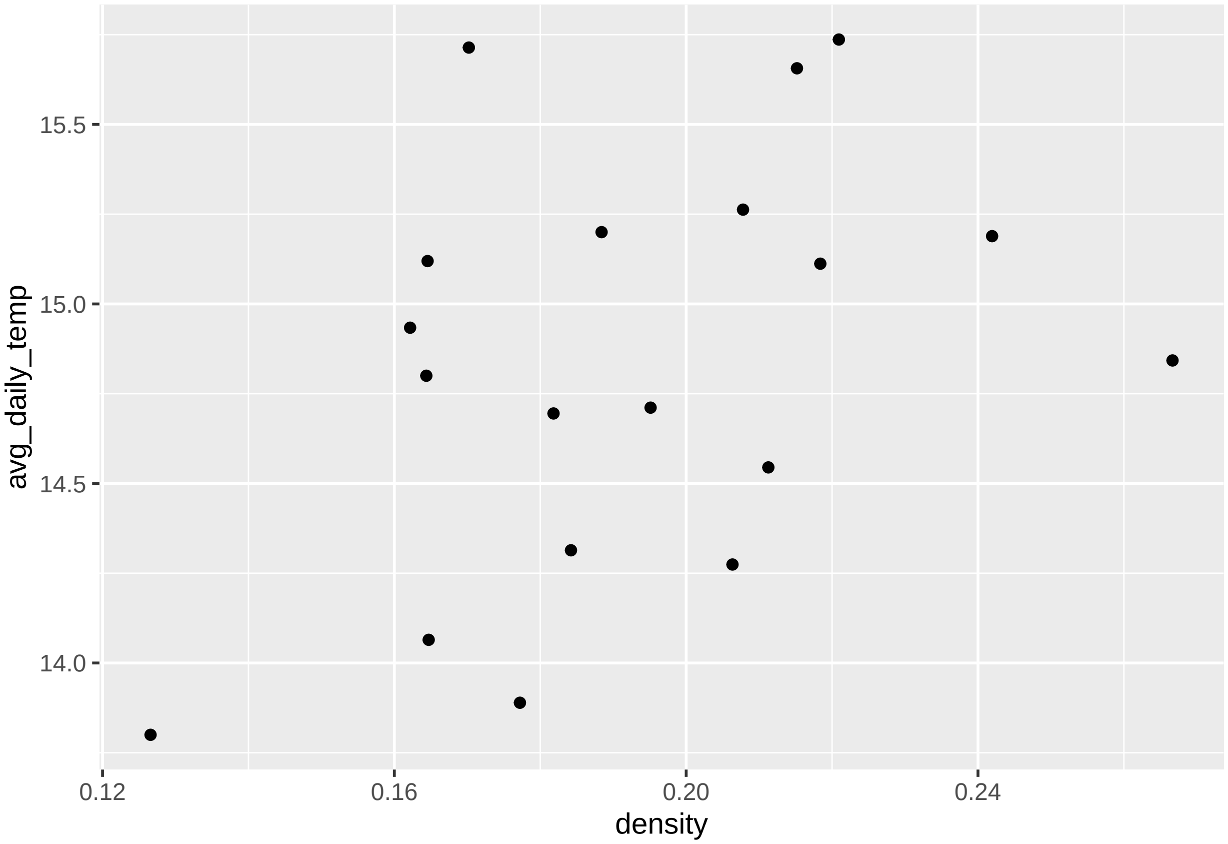
Are those columns continuous or categorical? What plot type have we talked about that might help us find a relationship between density and each of these variables (one at a time…)?

In small groups, make plots using the `sick_trout` data to try to figure out which environmental factor is causing problems in the trout.

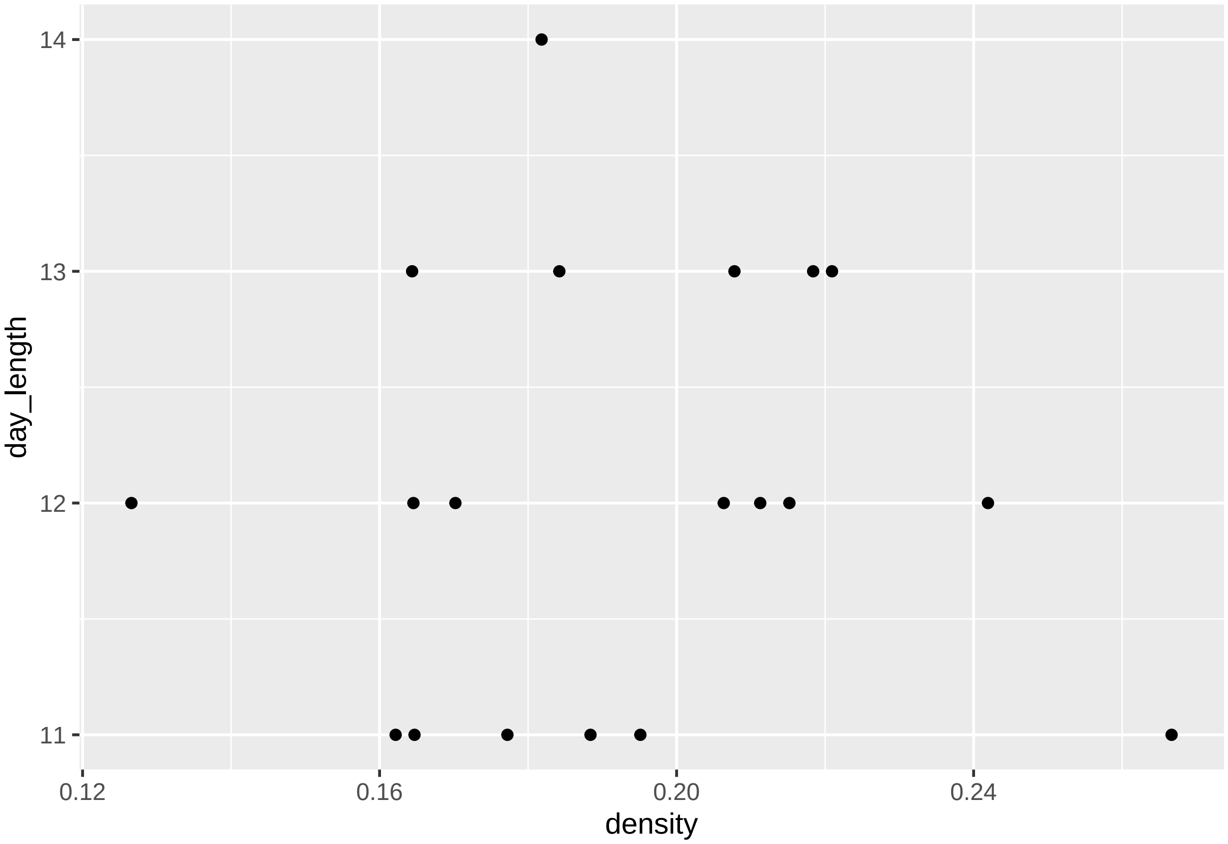
```
ggplot(sick_trout, aes(x = density, y = num_fish)) +
  geom_point()
```



```
ggplot(sick_trout, aes(x = density, y = avg_daily_temp)) +
  geom_point()
```



```
ggplot(sick_trout, aes(density, day_length)) +
  geom_point()
```



What do we think is the environmental driver causing issues with the trout?