

Code-a-long-2.3

What is behind the sick fish?

Goals

1. Students will apply coding techniques to find relationships between variables.
2. Students will generate data visualizations to help support or oppose an argument.

We know that there are tanks whose temperature are below the critical threshold for the immune systems of the fish species we are farming. However, there could be other factors contributing to the numbers of sick fish. After our class brainstormed more factors, the ichthyologists (fish scientists) measured: oxygen concentration and ammonia concentration (a proxy for waste buildup). We are going to look at these factors as well, to ensure we can address all of the factors affecting the fish health.

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

```
— Attaching core tidyverse packages — tidyverse 2.0.0 —
✓ dplyr      1.1.2      ✓ readr      2.1.4
✓ forcats    1.0.0      ✓ stringr    1.5.0
✓ ggplot2    3.4.2      ✓ tibble     3.2.1
✓ lubridate  1.9.2      ✓ tidyr      1.3.0
✓ purrr      1.0.2

— Conflicts — tidyverse_conflicts() —
* dplyr::filter() masks stats::filter()
* dplyr::lag()     masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
# read in the data, sick-fish.csv

data=read_csv("sick-fish.csv")
```

```
Rows: 1000 Columns: 13
— Column specification —
Delimiter: ","
chr  (1): species
dbl (11): tank_id, avg_daily_temp, num_fish, day_length, tank_volume, size_d...
lgl  (1): below

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.
```

```
# look at the data
```

```
view(data)
```

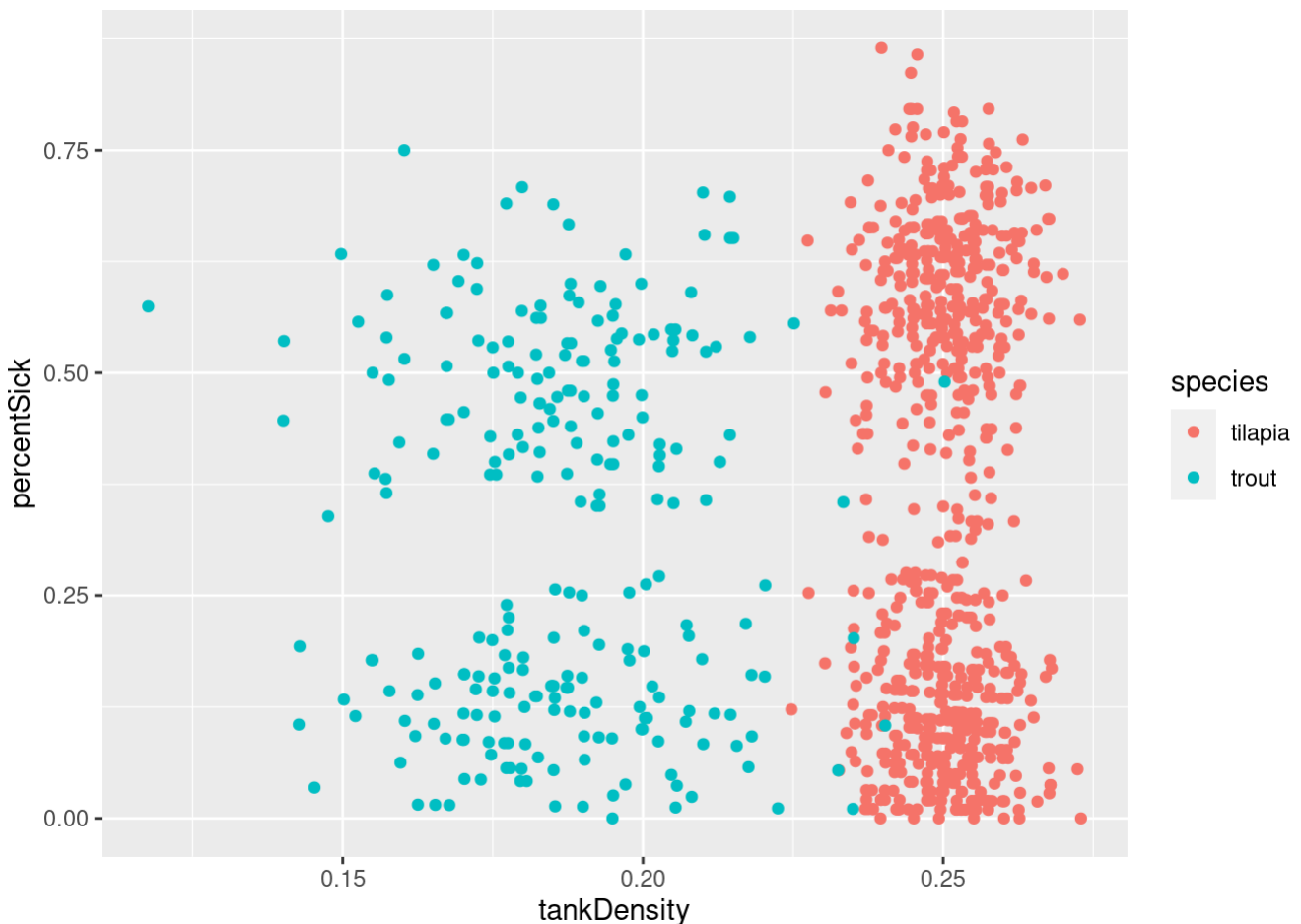
Our ichthyologist friends told us that density often contributes to the spread of any disease present in a system. We want to look at how density relates to the number of sick fish. Because we are in Antarctica, and obtaining supplies is quite difficult, not all of our tanks are from the same manufacturer and shipment. We have tanks of many different sizes. We know the size of each tank and the number of fish, so we can calculate the density. (Density = number / volume).

Create a variable in the data set for the density of fish per tank. Create a variable in the data set for the percentage of sick fish per tank.

```
data=data |>
  mutate(tankDensity=num_fish/tank_volume,
         percentSick=num_sick/num_fish)
```

Create a scatter plot to examine the relationship between density and the percentage of sick fish. In comments, explain why we are looking at the relationship between the density and the number of sick fish in a tank instead of the total number of fish in a given tank and the percentage of sick fish.

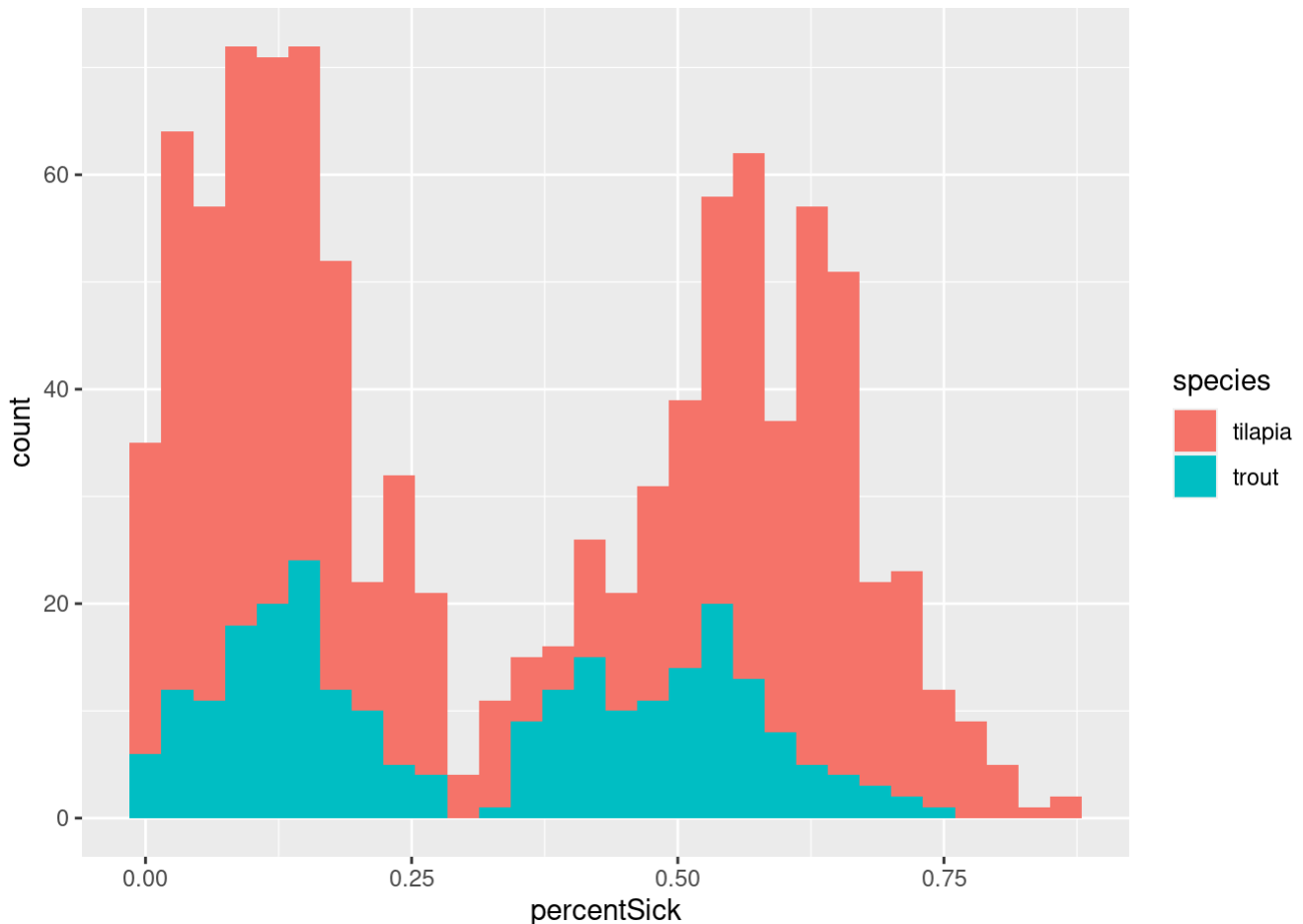
```
ggplot(data, mapping=aes(x=tankDensity, y=percentSick, color=species))+
  geom_point()
```



Let's look at a histogram of the `percentSick` variable, by species:

```
ggplot(data, mapping=aes(x=percentSick, fill=species))+  
  geom_histogram()
```

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.



At this point in class, student groups selected one of two fish species (tilapia or trout) and one of three variables (temperature, oxygen concentration, or ammonia concentration). Groups then created data subset to focus on their species and variable, and generated histograms and scatter plots. The goal was to determine if there was a visual pattern between their variable and the `percentSick` variable calculated above. Student groups then added their findings to a shared Google Slide file, and took turns presenting their findings.

Groups examining temperature created an additional bar plot using the `below` column in the data set. This is a categorical variable indicating whether the average temperature was above or below the critical threshold for fish immune systems. If TRUE, then that tank is below the critical threshold. If FALSE, then that tank is above the critical threshold. Create a histogram that examines sick fish and the tank temperature as this categorical variable.

Let's go through each combination. We can probably do a fair amount of copy/paste, and edit the specifics.

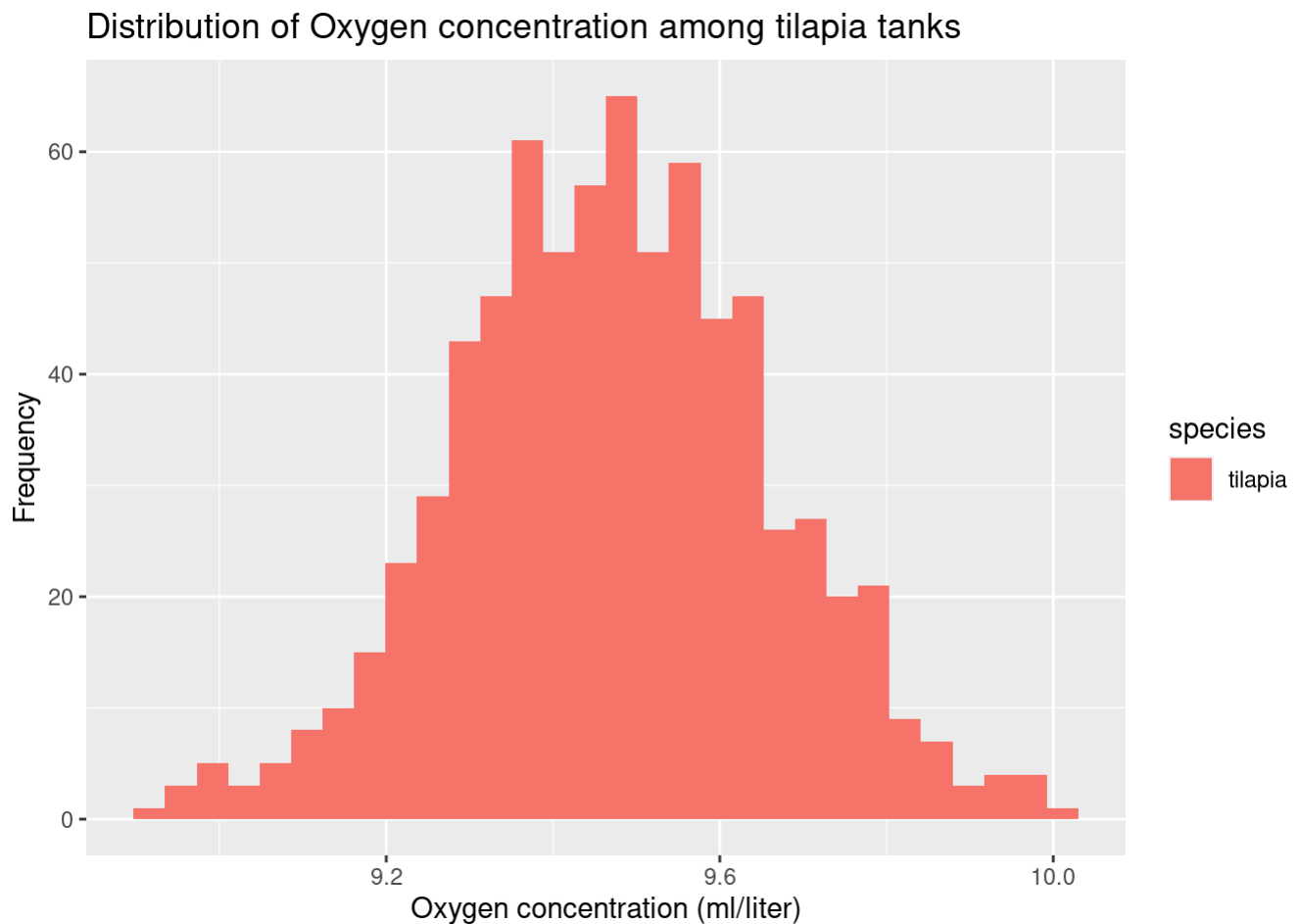
Tilapia and Oxygen

- create subset of data
- create histogram of O₂ concentration frequency across tanks
- create scatter plot of percent sick fish by O₂ concentration

```
tilapia=data |>
  filter(species=="tilapia")

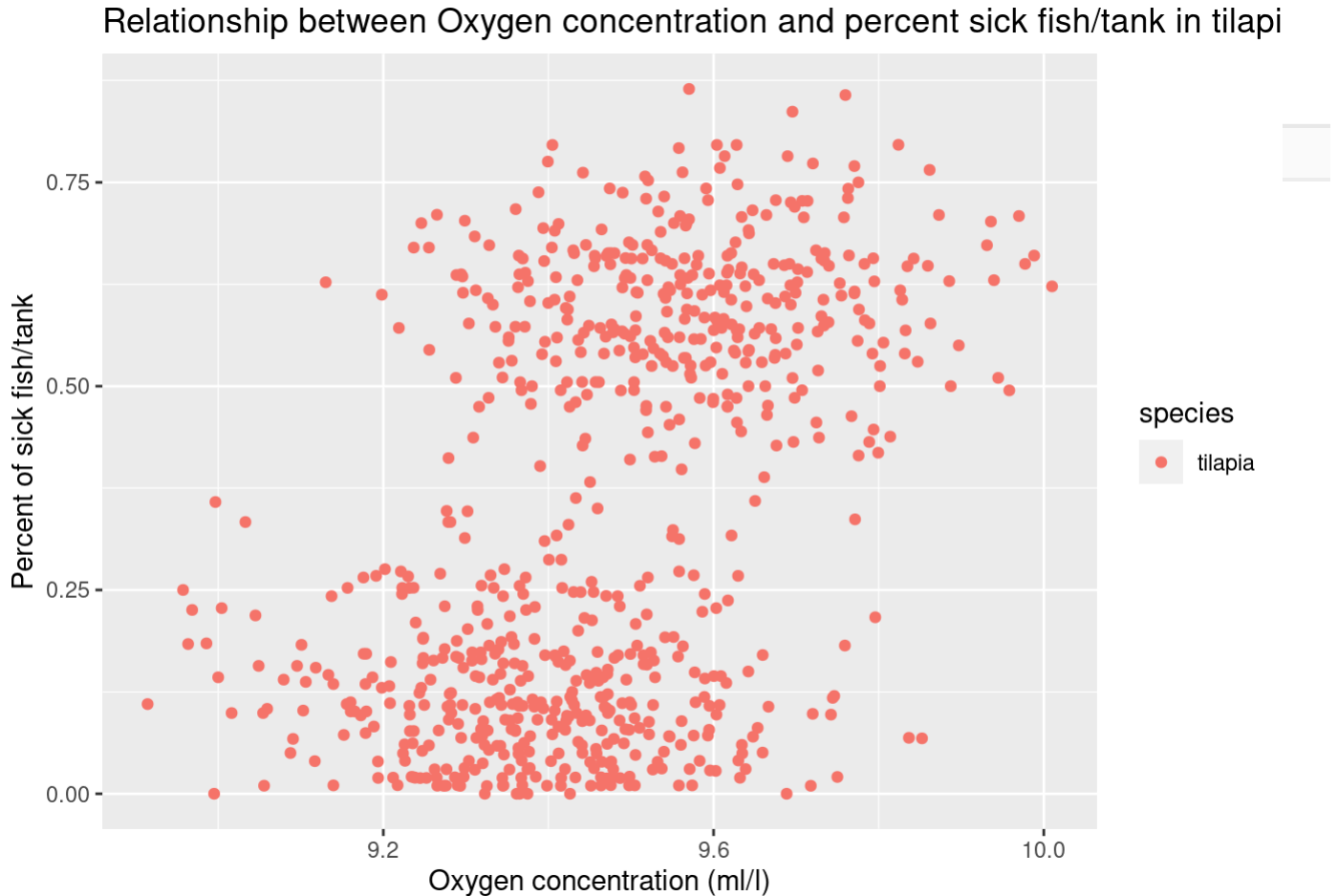
ggplot(tilapia, mapping=aes(x=oxygen, fill=species))+
  geom_histogram()+
  labs(x="Oxygen concentration (ml/liter)",
       y="Frequency",
       title="Distribution of Oxygen concentration among tilapia tanks")
```

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.



```
ggplot(tilapia, mapping=aes(x=oxygen, y=percentSick, color=species))+
  geom_point()+
  labs(x="Oxygen concentration (ml/l)",
```

```
y="Percent of sick fish/tank",
title="Relationship between Oxygen concentration and percent sick fish/tank"
```



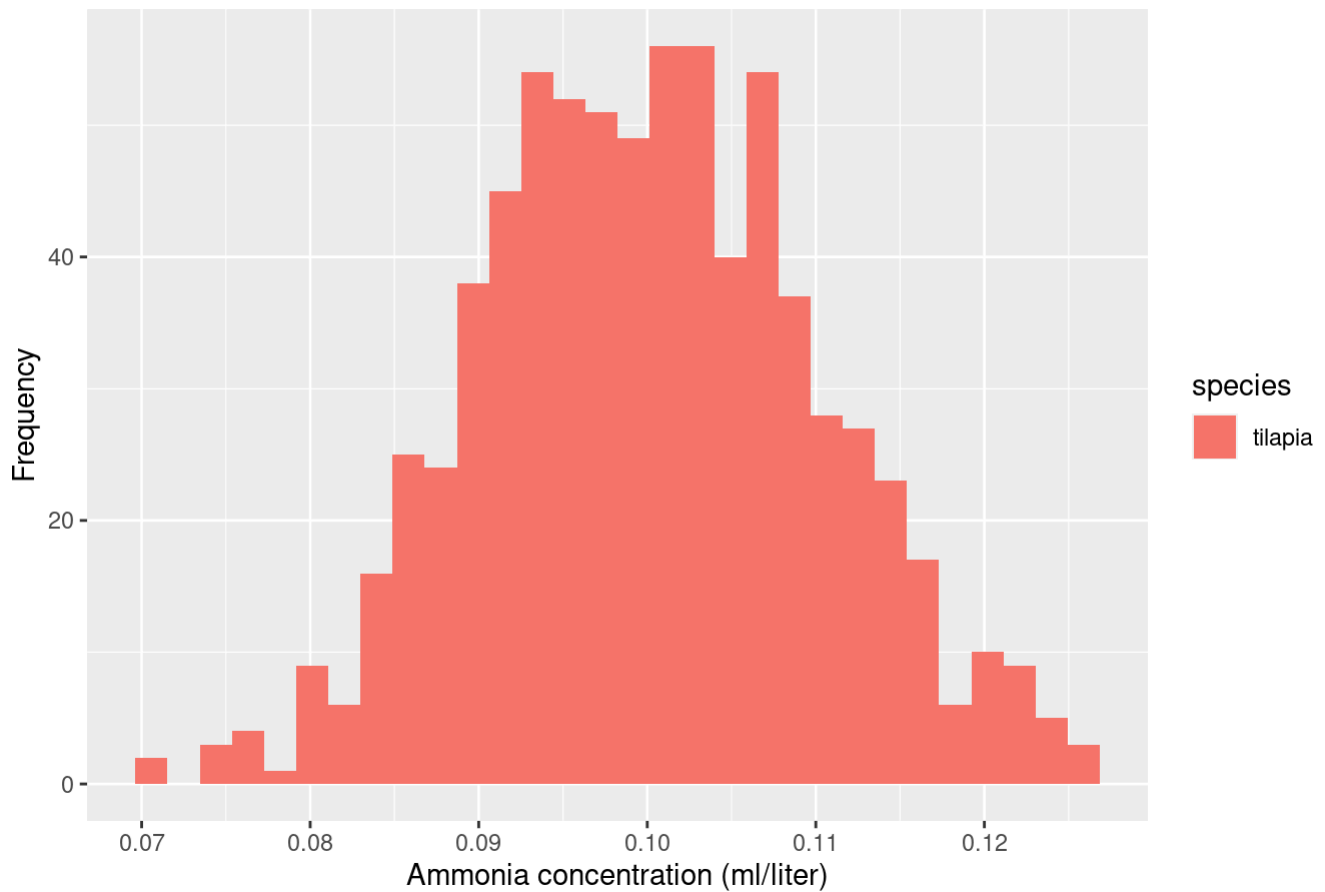
Tilapia and Ammonia

- create subset of data
- create histogram of NH_3 concentration frequency across tanks
- create scatter plot of percent sick fish by NH_3 concentration

```
ggplot(tilapia, mapping=aes(x=ammonia, fill=species))+
  geom_histogram()+
  labs(x="Ammonia concentration (ml/liter)",
       y="Frequency",
       title="Distribution of Ammonia concentration among tilapia tanks")
```

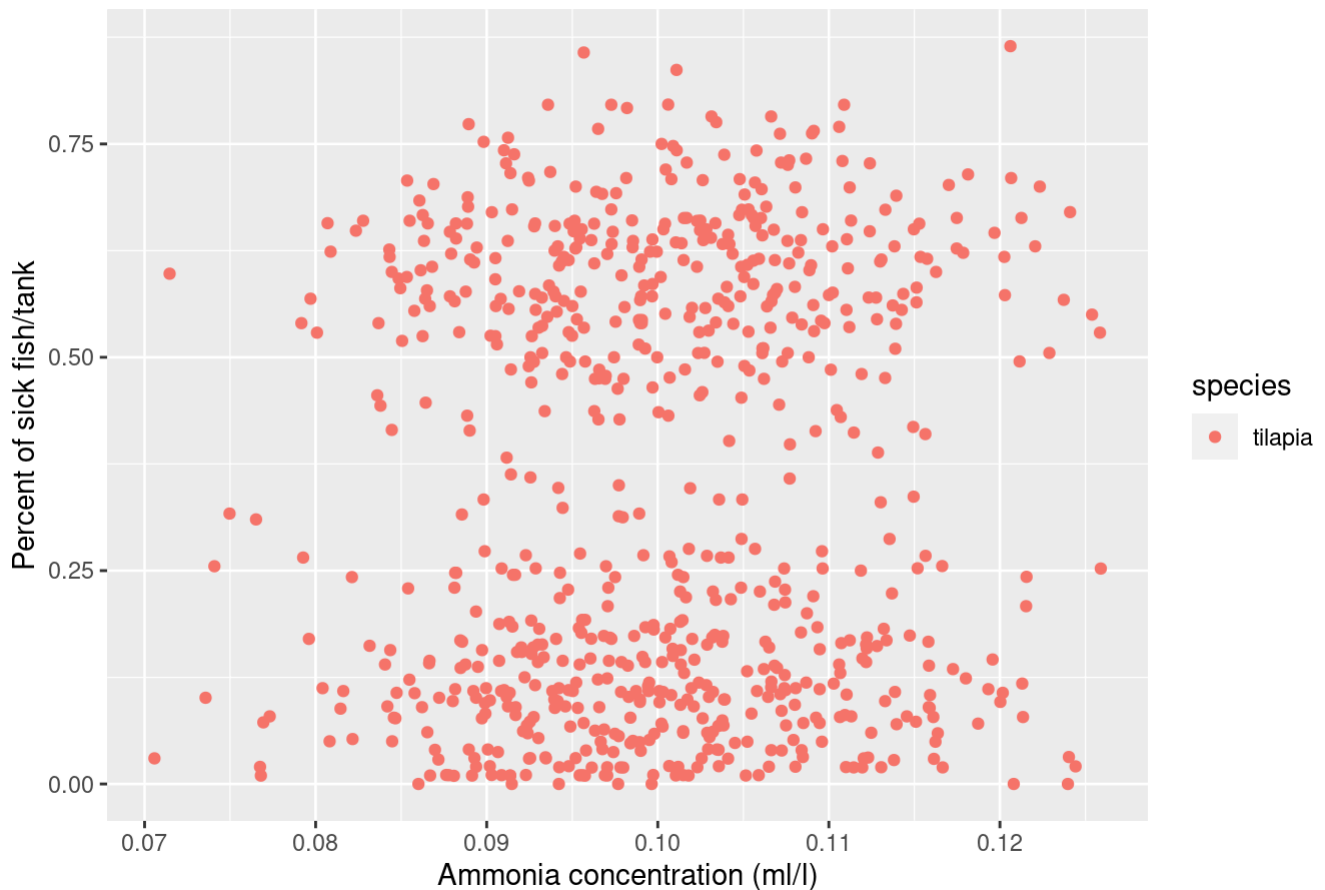
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

Distribution of Ammonia concentration among tilapia tanks



```
ggplot(tilapia, mapping=aes(x=ammonia, y=percentSick, color=species))+  
  geom_point()+  
  labs(x="Ammonia concentration (ml/l)",  
        y="Percent of sick fish/tank",  
        title="Relationship between Ammonia concentration and percent sick fish/ta
```

Relationship between Ammonia concentration and percent sick fish/tank in tilapia



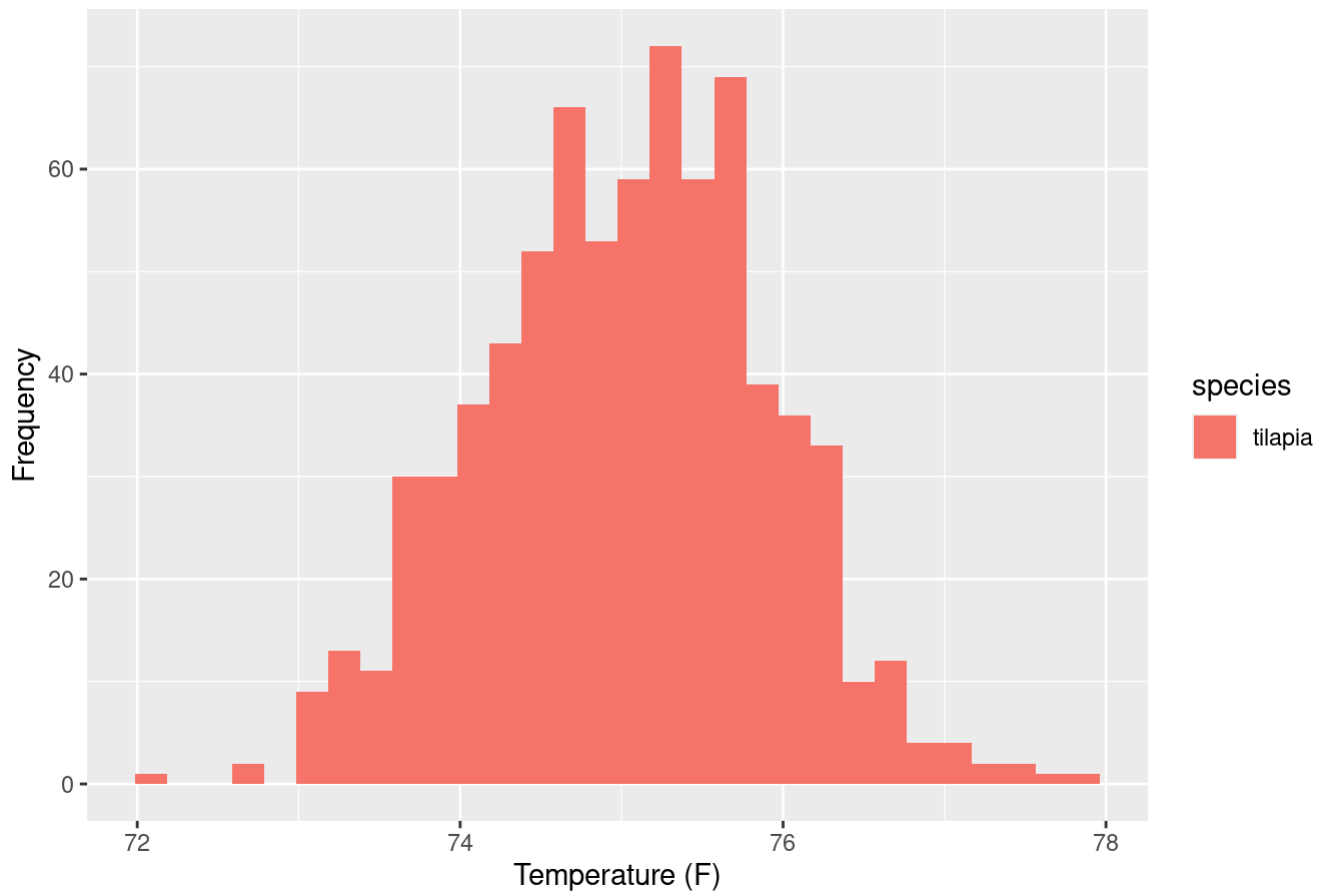
Tilapia and Temperature

- create subset of data
- create histogram of temperature frequency across tanks
- create scatter plot of percent sick fish by temperature
- create a bar plot comparing the mean temperature, grouping by the **below** column (this may require performing an additional data transformation)

```
ggplot(tilapia, mapping=aes(x=avg_daily_temp_F, fill=species))+  
  geom_histogram()+  
  labs(x="Temperature (F)",  
        y="Frequency",  
        title="Distribution of average daily temperature among tilapia tanks")
```

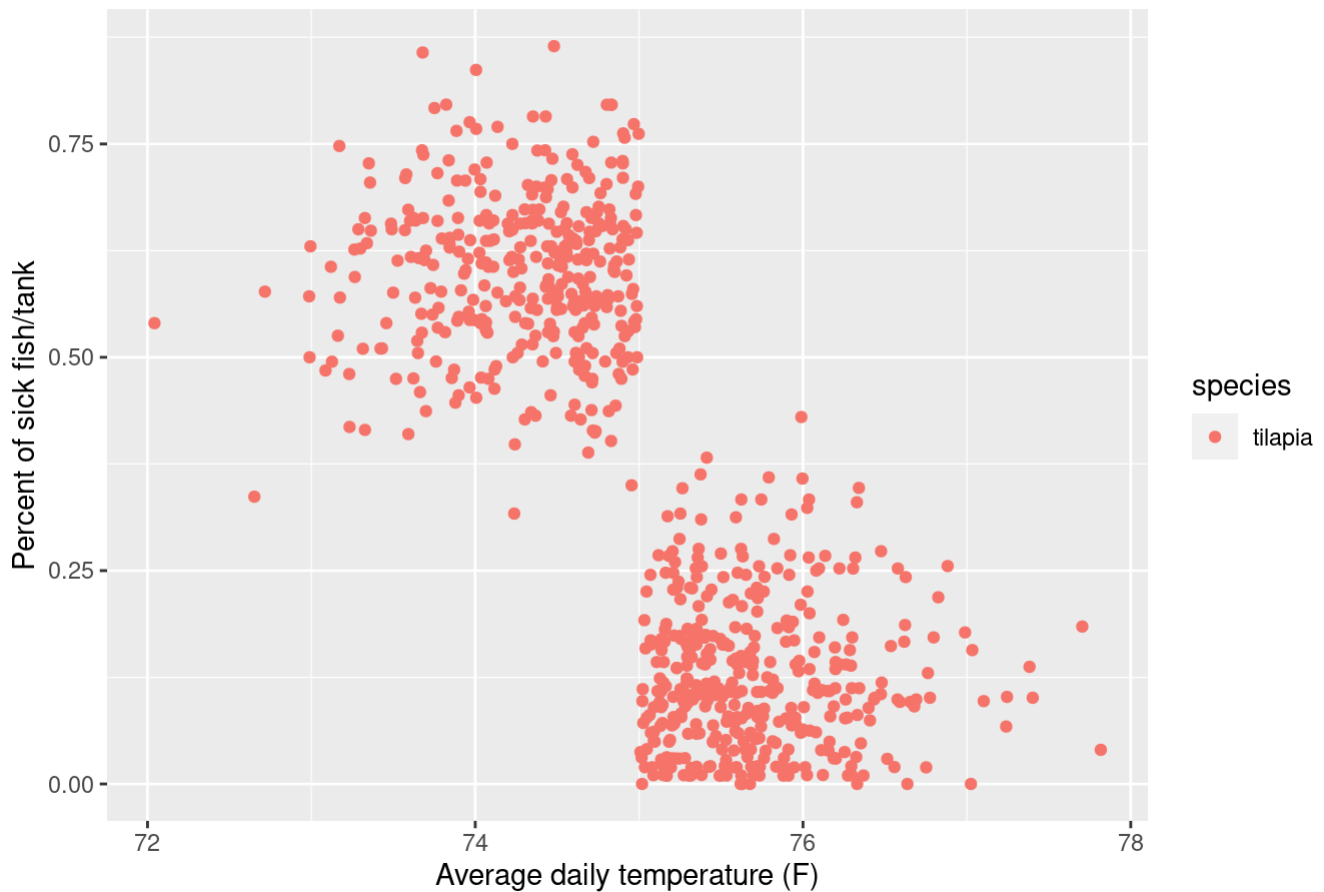
``stat_bin()`` using ``bins = 30``. Pick better value with ``binwidth``.

Distribution of average daily temperature among tilapia tanks



```
ggplot(tilapia, mapping=aes(x=avg_daily_temp_F, y=percentSick, color=species))+  
  geom_point()+  
  labs(x="Average daily temperature (F)",  
        y="Percent of sick fish/tank",  
        title="Relationship between average daily temperature and percent sick fish/tank")
```


Relationship between average daily temperature and percent sick fish/tank in t



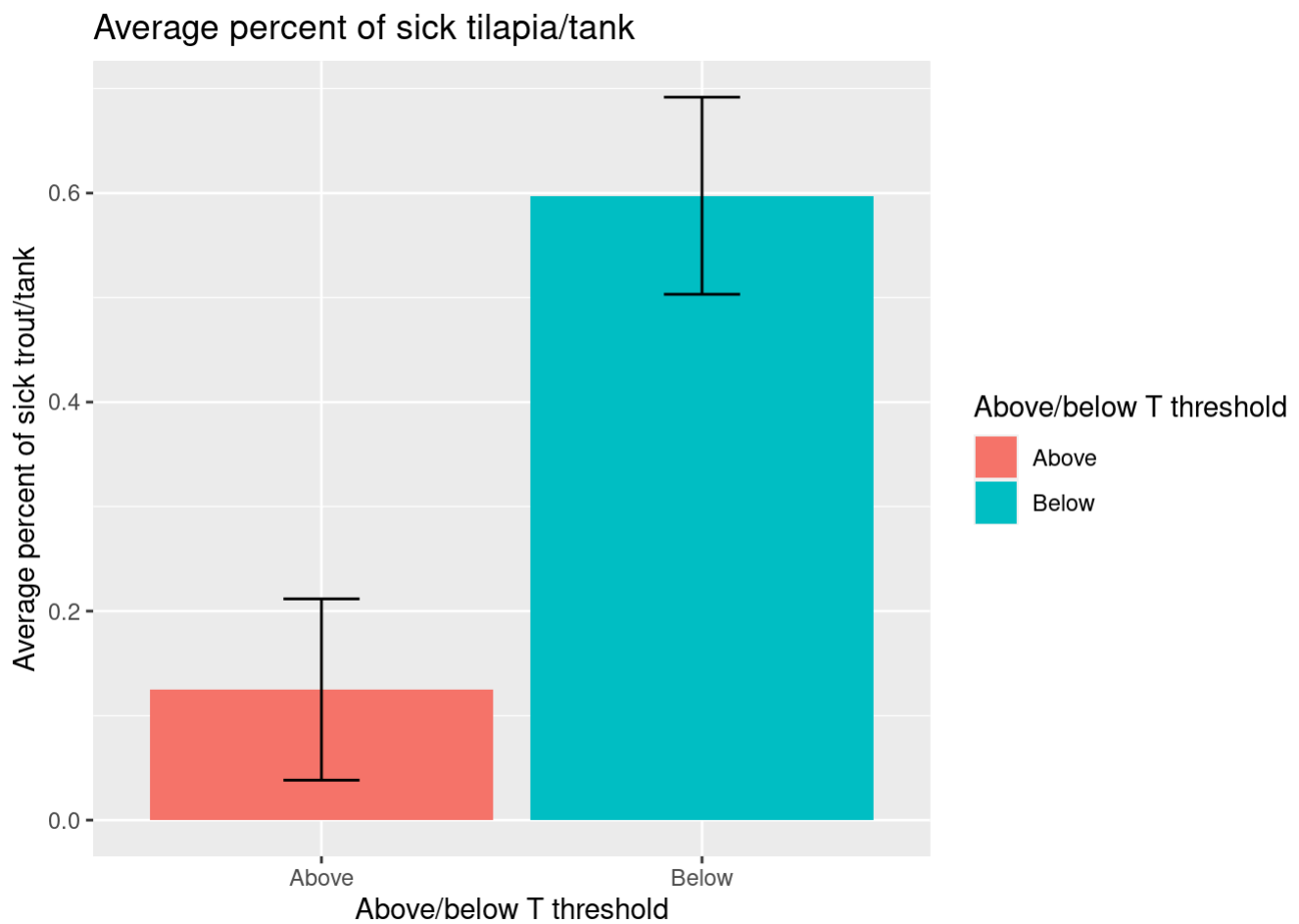
```
# with a bar plot
tilapiaSummaryData <- tilapia |>
  group_by(below) |>
  summarize(meanSick = mean(percentSick), sdSick = sd(percentSick))

view(tilapiaSummaryData)

ggplot(tilapiaSummaryData, aes(x = below, y = meanSick, fill = below)) +
  geom_bar(stat = "identity") +
  geom_errorbar(aes(ymin = meanSick - sdSick, ymax = meanSick + sdSick, width = 0.5)) +
  labs(title="Average percent of sick tilapia/tank",
       fill="Above/below T threshold",
       x="Above/below T threshold",
       y="Average percent of sick trout/tank")+
  scale_x_discrete(labels = c("FALSE" = "Above", "TRUE" = "Below"))+
  scale_x_discrete(labels = c("FALSE" = "Above", "TRUE" = "Below"))+
  scale_fill_discrete(labels = c("FALSE" = "Above", "TRUE" = "Below"))
```

Scale for x is already present.

Adding another scale for x, which will replace the existing scale.



Trout and Oxygen

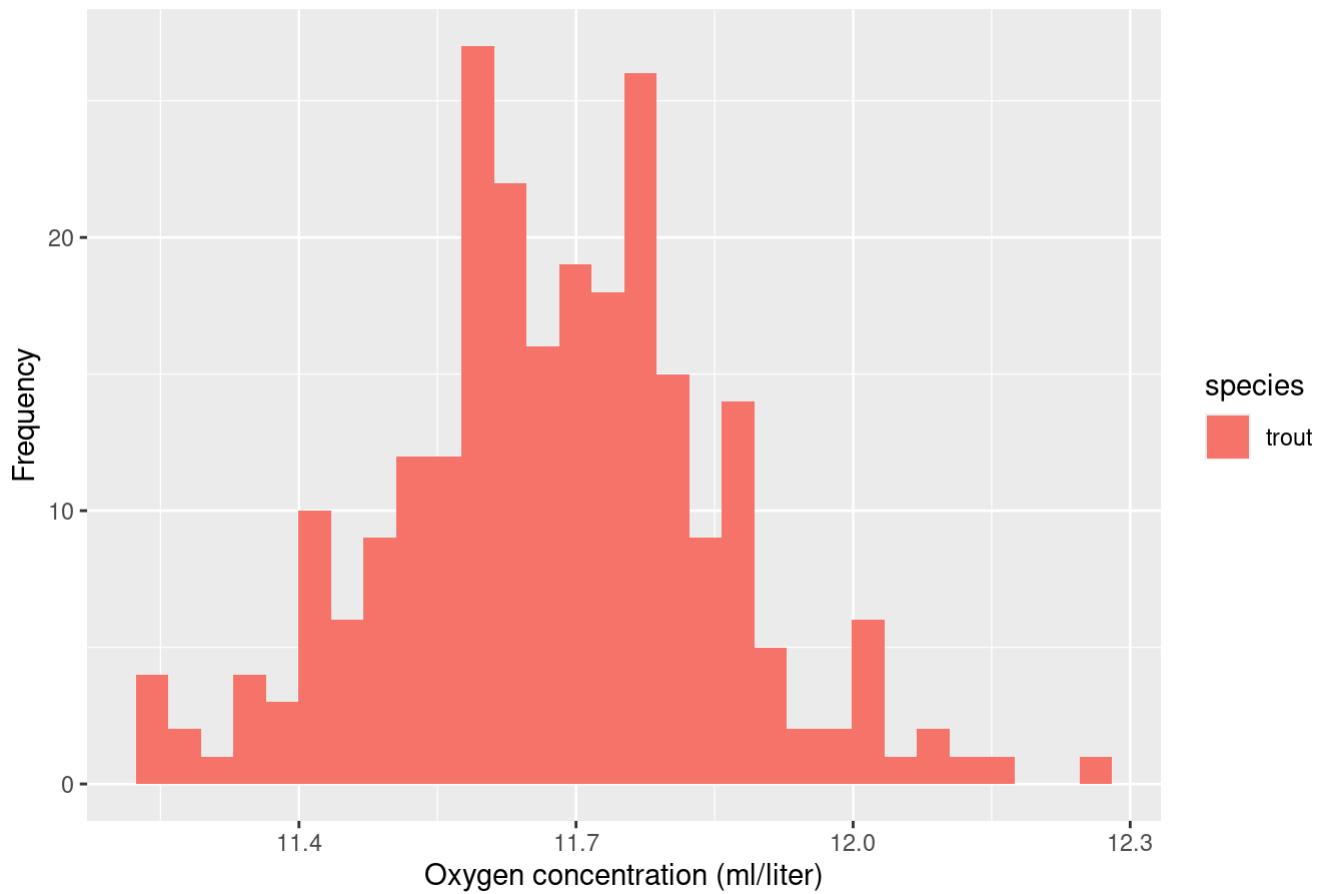
- create subset of data
- create histogram of O₂ concentration frequency across tanks
- create scatter plot of percent sick fish by O₂ concentration

```
trout=data |>
  filter(species=="trout")

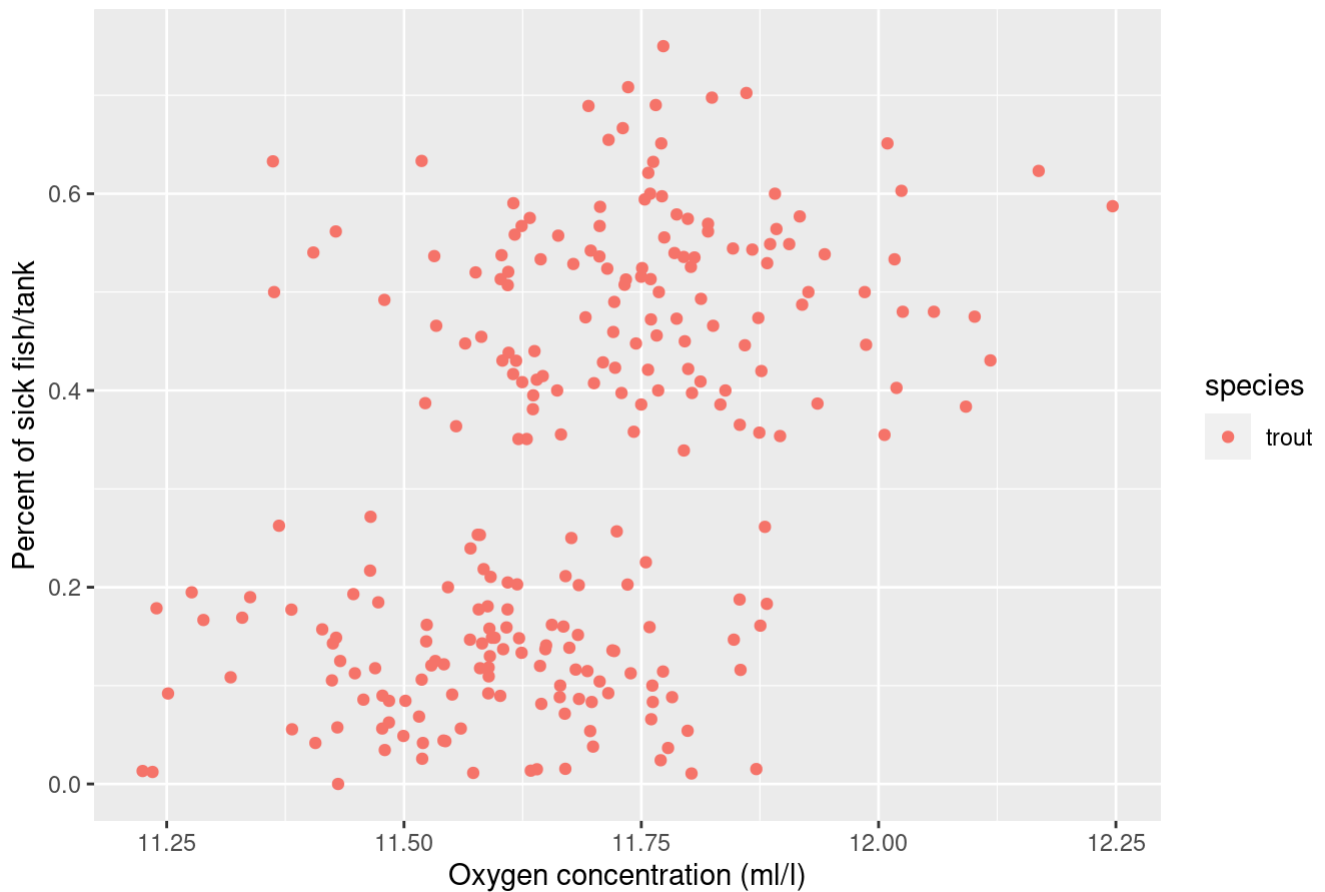
ggplot(trout, mapping=aes(x=oxygen, fill=species))+
  geom_histogram()+
  labs(x="Oxygen concentration (ml/liter)",
       y="Frequency",
       title="Distribution of Oxygen concentration among trout tanks")
```

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

Distribution of Oxygen concentration among trout tanks



Relationship between Oxygen concentration and percent sick fish/tank in trout



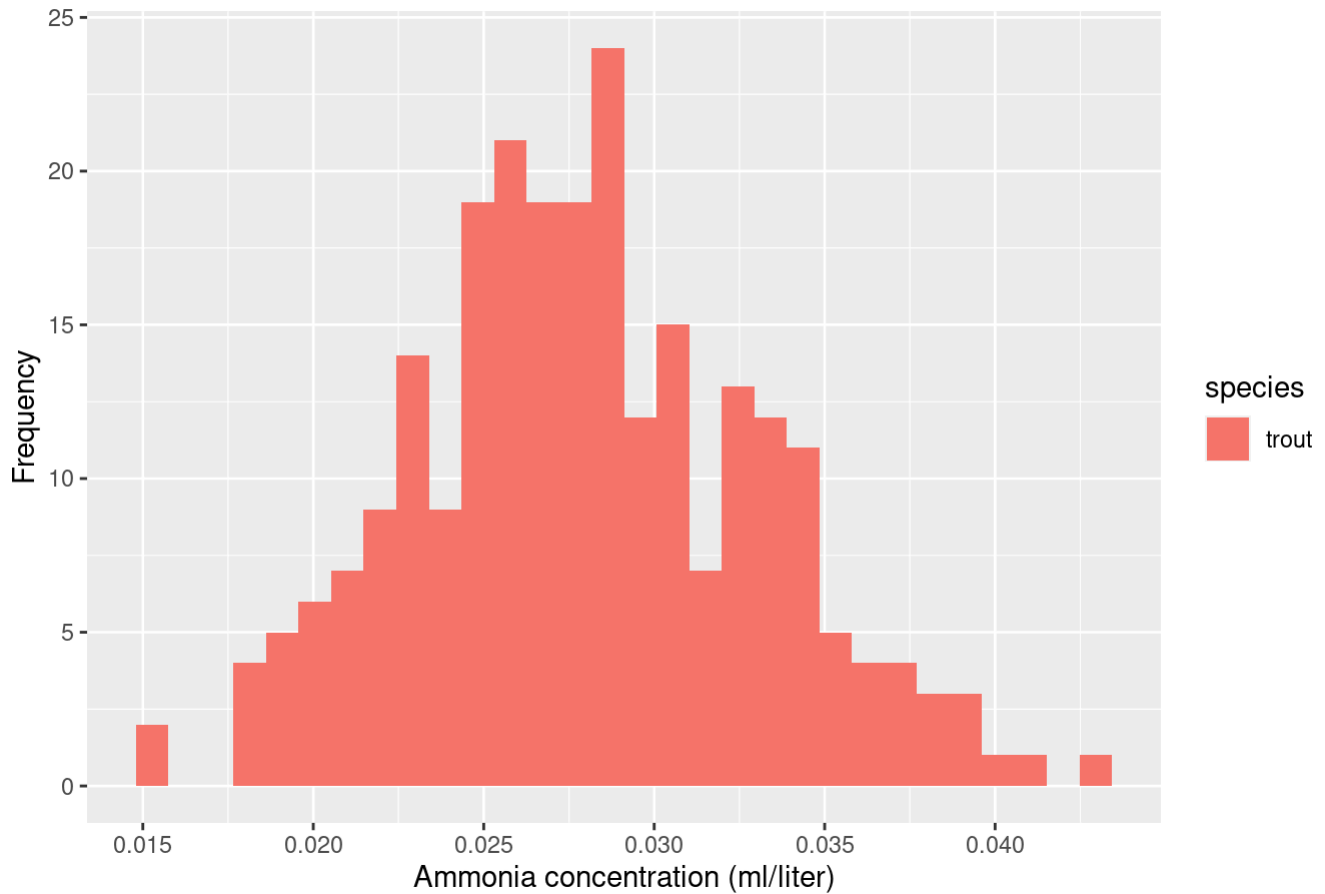
Trout and Ammonia

- create subset of data
- create histogram of NH_3 concentration frequency across tanks
- create scatter plot of percent sick fish by NH_3 concentration

```
ggplot(trout, mapping=aes(x=ammonia, fill=species))+  
  geom_histogram()+  
  labs(x="Ammonia concentration (ml/liter)",  
        y="Frequency",  
        title="Distribution of Ammonia concentration among trout tanks")
```

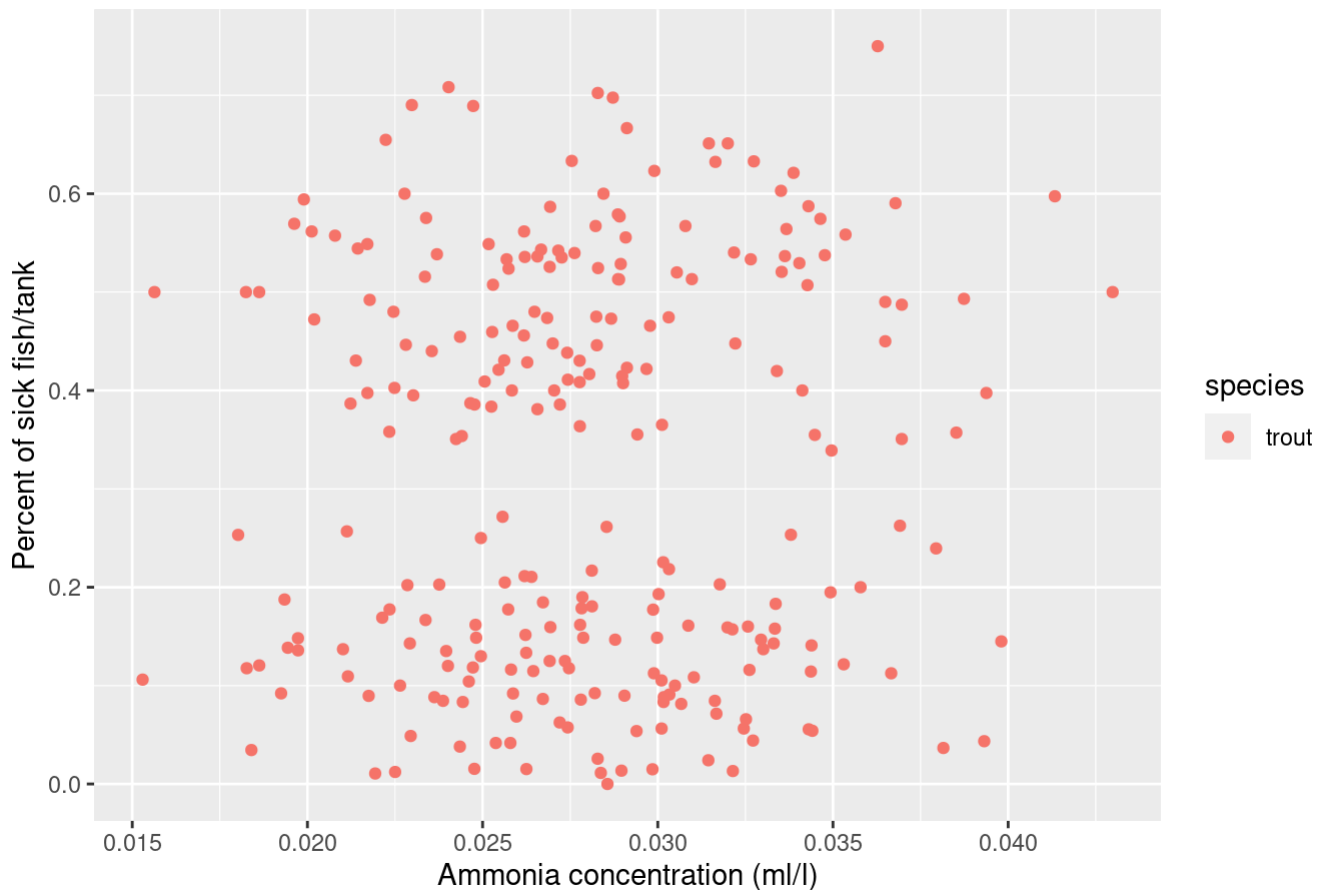
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

Distribution of Ammonia concentration among trout tanks



```
ggplot(trout, mapping=aes(x=ammonia, y=percentSick, color=species))+  
  geom_point()+  
  labs(x="Ammonia concentration (ml/l)",  
        y="Percent of sick fish/tank",  
        title="Relationship between Ammonia concentration and percent sick fish/tank")
```

Relationship between Ammonia concentration and percent sick fish/tank in trout



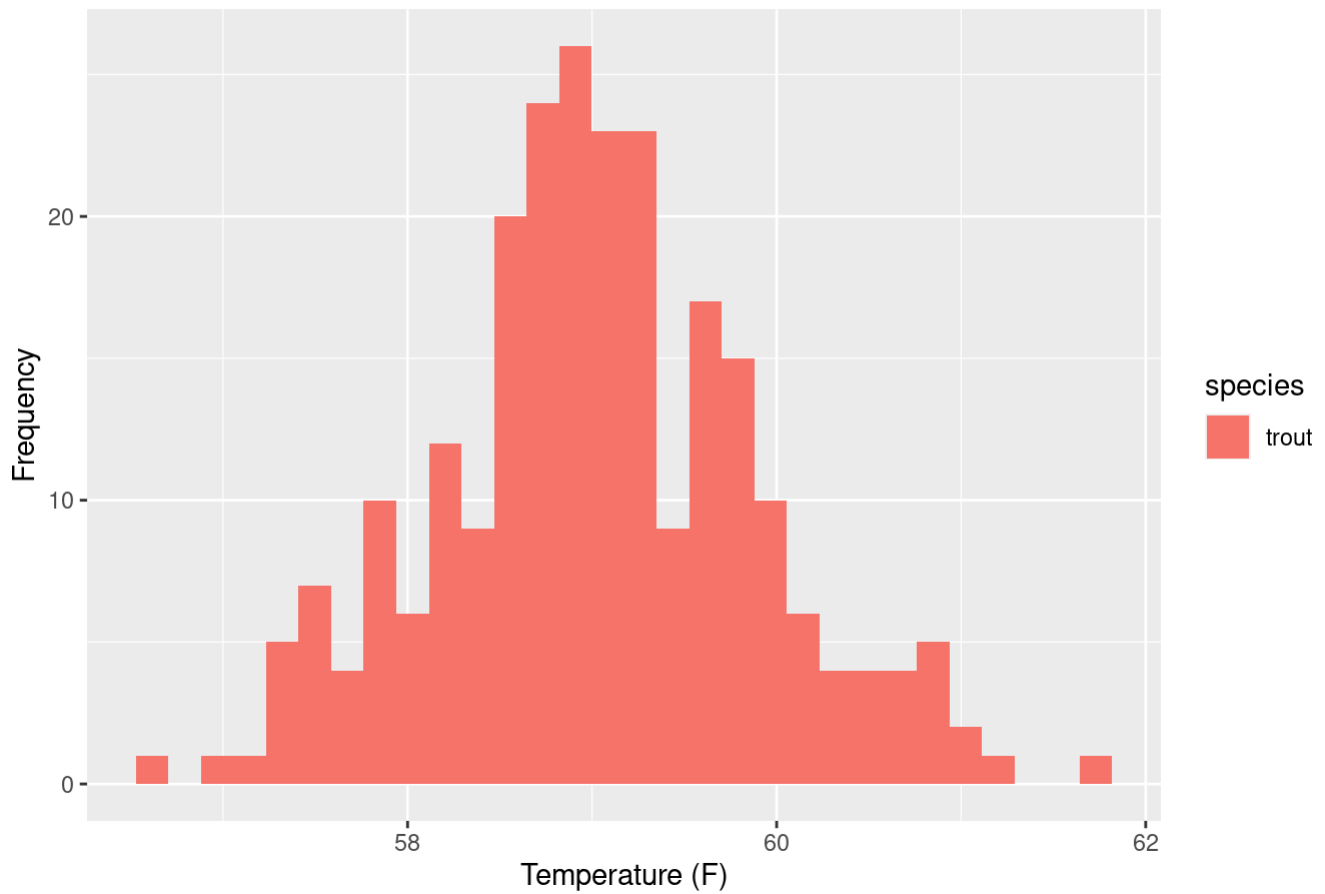
Trout and Temperature

- create subset of data
- create histogram of temperature frequency across tanks
- create scatter plot of percent sick fish by temperature
- create a bar plot comparing the mean temperature, grouping by the `below` column (this may require performing an additional data transformation)

```
ggplot(trout, mapping=aes(x=avg_daily_temp_F, fill=species))+  
  geom_histogram()+  
  labs(x="Temperature (F)",  
        y="Frequency",  
        title="Distribution of average daily temperature among trout tanks")
```

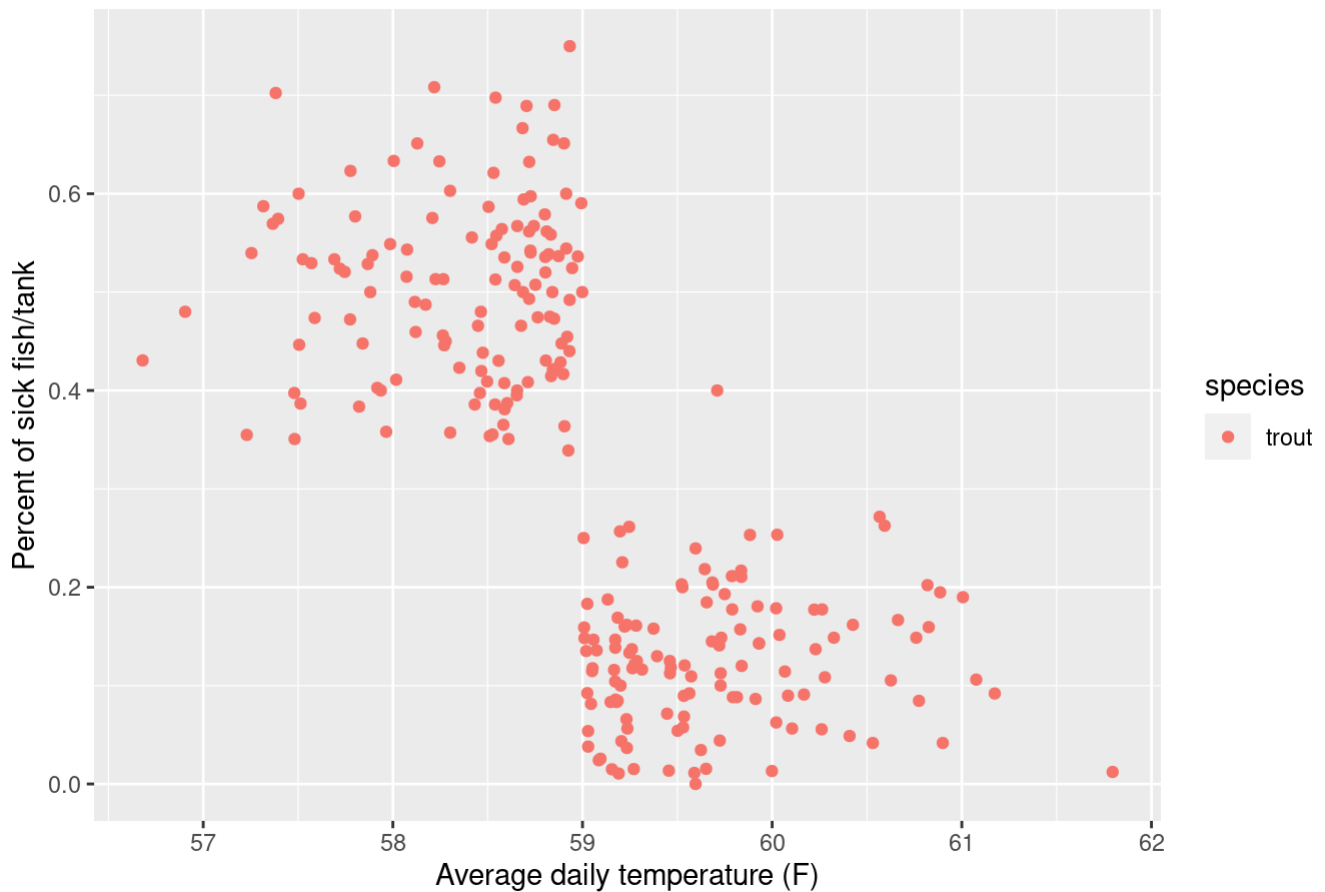
``stat_bin()`` using ``bins = 30``. Pick better value with ``binwidth``.

Distribution of average daily temperature among trout tanks



```
ggplot(trout, mapping=aes(x=avg_daily_temp_F, y=percentSick, color=species))+  
  geom_point()+  
  labs(x="Average daily temperature (F)",  
        y="Percent of sick fish/tank",  
        title="Relationship between average daily temperature and percent sick fish/tank")
```

Relationship between average daily temperature and percent sick fish/tank in tr



```
# with a bar plot
troutSummaryData <- trout |>
  group_by(below) |>
  summarize(meanSick = mean(percentSick), sdSick = sd(percentSick))

#view(troutSummaryData)

ggplot(troutSummaryData, aes(x = below, y = meanSick, fill = below)) +
  geom_bar(stat = "identity") +
  geom_errorbar(aes(ymin = meanSick - sdSick, ymax = meanSick + sdSick, width = 0.5)) +
  labs(title="Average percent of sick trout/tank",
       fill="Above/below T threshold",
       x="Above/below T threshold",
       y="Average percent of sick trout/tank")+
  scale_x_discrete(labels = c("FALSE" = "Above", "TRUE" = "Below"))+
  scale_fill_discrete(labels = c("FALSE" = "Above", "TRUE" = "Below"))
```


Average percent of sick trout/tank

