Module 2: Good Food Gone Bad

Keaton Wilson, modified by EKB

Descriptive Statistics and Data Visualization

Student Learning Outcomes

1. Students will be able to apply basic data science knowledge to find the cause of a real-world scenario–food poisoning!
2. Students will be able to generate two types of plots using base R syntax to visualize a single continuous variable (histograms and distributions) and two continuous variables (scatter plots).
3. Students will be able to use visual-thinking skills to create visualizations that allow them to explore patterns in data, draw inferences, and create solutions.

Introduction to the problem

We have a wave of people getting sick across the team. People are coming in complaining of stomach sickness. Doctors have ruled out a communicable viral infection like norovirus, so it seems likely to be a food contamination issue.

The two main sources of food that are grown on site and distributed to team members are plants grown in hydroponic greenhouses (mostly Swiss chard, cucumbers and radishes) and fish (tilapia, a tolerant warm-water species, and rainbow trout, a cold-water species). The combination of aquaculture and hydroponics is called **aquaponics**.

Team members' diets vary in composition; people are allowed to choose how much of different food sources they eat.

Fortunately, we have some data we can use to investigate! We have data on the following:

- which team members are sick and how many times they've gone to the doctor
- some information about each team member, such as:
  - sex, age, height, occupation

- how much fish and/or plant material they incorporate into their diets

The Data

First we're going to pull in the data and give it a quick inspection/exploration before we start.

As usual, we start by calling the `tidyverse`.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.2
## v ggplot2   4.0.0     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

To bring our data into R, we use a function called `read_csv()`. CSV (comma-separated values) are efficient ways to save 2-dimensional ("spreadsheet") data.

```
sick <- read_csv("../data/sick_data.csv")
```

```
## Rows: 349 Columns: 10
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): last, first, sex, specialties
## dbl (6): age, height_cm, weight_kg, perc_fish, perc_plant, doctor_trips
##
## 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(sick)
```

```
## # A tibble: 6 x 10
##   last    first sex     age height_cm weight_kg specialties perc_fish perc_plant
##   <chr>   <chr> <chr> <dbl>     <dbl>     <dbl> <chr>           <dbl>      <dbl>
## 1 Gonzal~ Ange~ M        35      169.      51.4 Hydrology       0.994    0.00620
## 2 Navrat~ John  M        19      112.      96.3 Genetics        0.297    0.703
## 3 Duff    Josh~ M        26      133.      52.1 Horticultu~     0.514    0.486
## 4 Dottson Juli~ M        36      140.      52.6 Climatology     0.686    0.314
## 5 al-Sul~ Mune~ M        26      194.      52.2 Geology         0.292    0.708
## 6 Galleg~ Rich~ M        29      153.      98.1 Climatology     0.329    0.671
## # i 1 more variable: doctor_trips <dbl>
```

Let's talk about the data we have:

- How many observations (rows) do we have?
- What type of information does a single row represent?
- How many variables (columns) do we have?

Which variables are we particularly interested in? Are they continuous or categorical? Does it matter?

Group Discussion

How might we figure out what is causing the problem? Try to focus on potential solutions that involve the data we already have.

Spend 5 minutes brainstorming in your groups how you might figure out whether plants or fish are the culprits? Be ready to report out.

Write a research question with a dependent variable and an independent variable (see the "Aquaponics & Research Questions" slide deck for more info).

Descriptive Statistics and Data Visualization

In order to begin addressing the question of what might be causing illness in our crew (and lots of other questions!), we want to start with *descriptive statistics* and *data visualization*.

In this course, we will be working with two types of statistics: *descriptive* and *inferential*.

- **Descriptive** statistics–also called **summary** statistics–are ways of presenting, organizing, and summarizing data. Data visualization is often associated with descriptive statistics
- **Inferential** statistics help us draw reasonable conclusions about a population based on the data we observed in a sample.

In Module 2, we will be focusing exclusively on descriptive statistics and data visualization techniques.

As a quick overview, descriptive statistics often include 3 elements:

- *distribution* of the data
- measures of *central tendency*: mean, median, and mode
- measures of *variation*: range and standard deviation

The slide deck with more information about descriptive statistics is linked on D2L.

For a nice overview of descriptive statistics, check out this website I showed in class. It goes a bit further than we go in this class, but it is a nice place to review!

Let's Practice

Now that we have a more formalized understanding of measures of central tendency and measures of variation, let's put them into use to learn a little bit more about our data. In small groups, calculate the mean (`mean()`) and standard deviation (`sd()`) for the percent fish in our sick crew's diets.

```
sick %>%
  summarize(mean_fish = mean(perc_fish, na.rm = TRUE),
            sd_fish = sd(perc_fish, na.rm = TRUE))
```

```
## # A tibble: 1 x 2
##   mean_fish sd_fish
##       <dbl>   <dbl>
## 1     0.537   0.250
```

Data Visualization in R

Combining data visualization with descriptive statistics is a great way to understand our data!

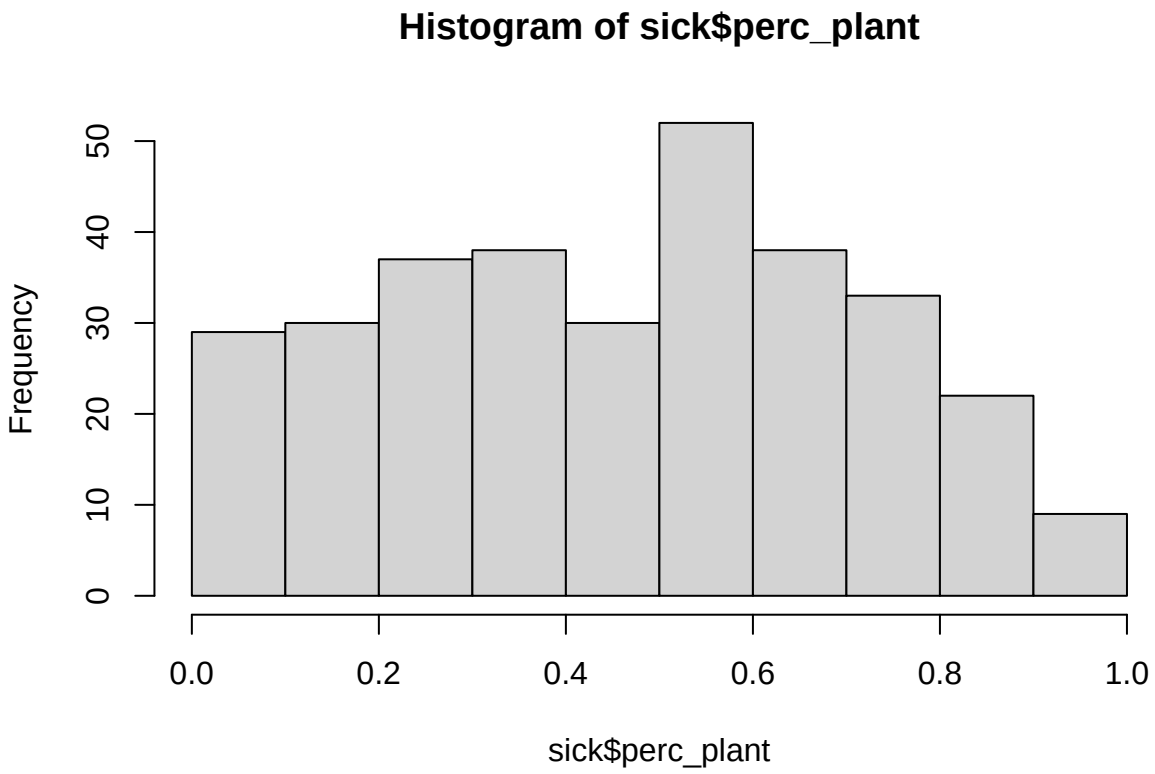
There are two main ways to make data visualizations in R: through base R, which is the syntax we learned in our first week of coding, and through `ggplot2`, which is a package in the `tidyverse`.

For the majority of the semester, we will be plotting in with `ggplot2`, which is a fun and powerful tool. However, to plot even a simple plot, `ggplot2` takes some explanation. We will talk about `ggplot2` in our *next class* but for today, let's first make some quick-and-dirty plots in *base R*.

Histograms

We've calculated some descriptive statistics about the percents of fish and plants in our sick crew members' diets, but it didn't tell us too much. Let's try some data visualization to see if that gives us any additional information.

```
# histogram in base R
hist(sick$perc_plant)
```



Group Discussion

Discuss this histogram with your group members. Some questions to consider:

- What *is* a histogram?
- What does a histogram tell us?
- What does each axis mean? (x-axis is horizontal, y-axis is vertical)
- What, if any, conclusions can we draw from this histogram in regards to the percent plants in diets and sickness?
- How can we improve this visualization?

Take about 5 minutes. Be ready to report out.

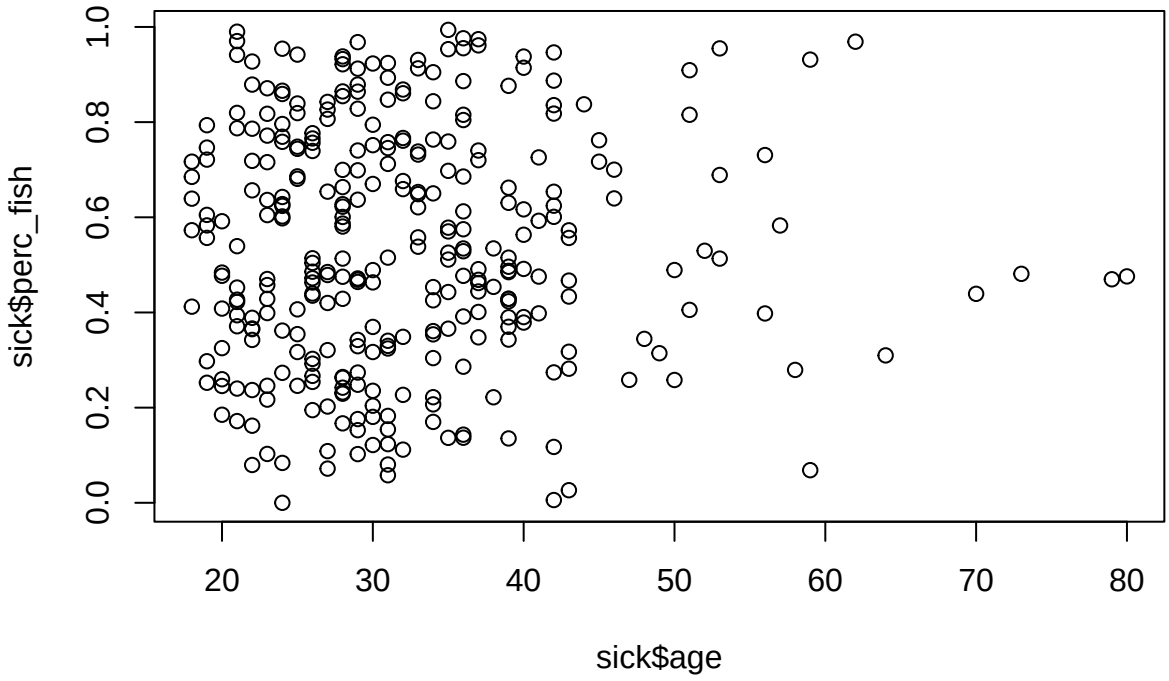
Group Brainstorm

We've covered one type of visualization that just shows one variable... but what we're really interested in is figuring out if fish or plants are the culprit in food poisoning. Spend about 5 minutes in your groups sketching out a visualization that might give us insight into this.

Scatter plots

Scatter plots allow us to visualize the relationship between two continuous (or numeric) variables. For example, we can use a scatter plot to see if there is a relationship between how old a crew member is and how much fish is in their diet.

```
# scatter plot in base R
plot(x = sick$age, y = sick$perc_fish)
```

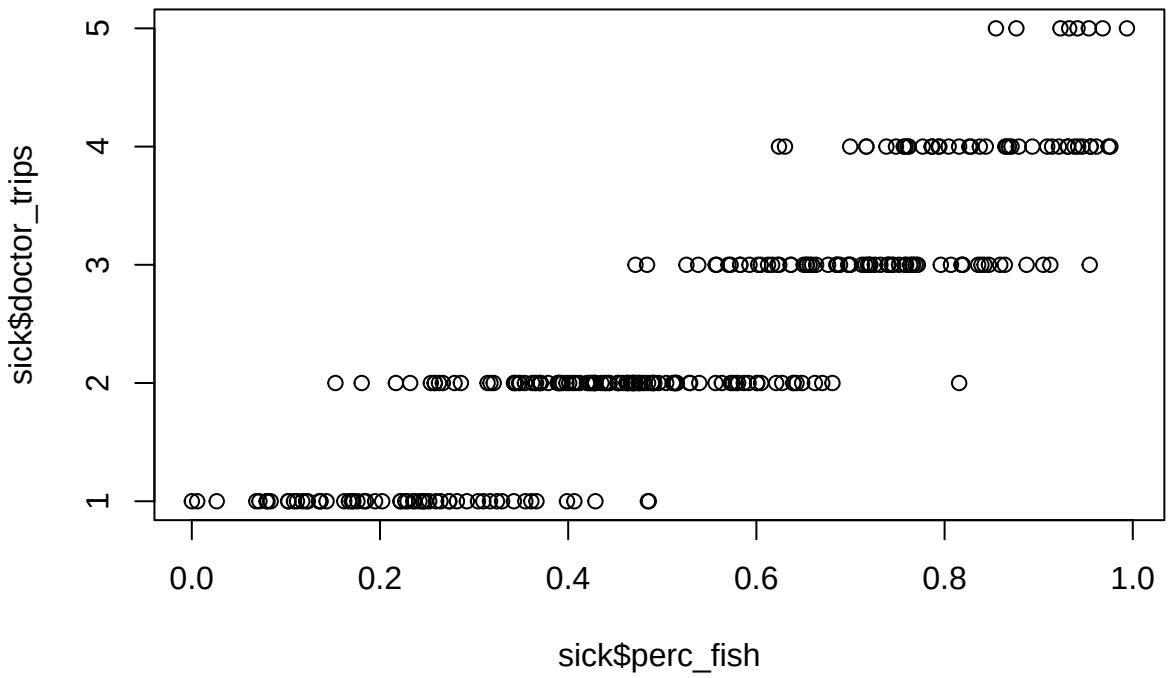


This isn't super informative, because there isn't really a relationship between a person's age and the amount of fish they consume. At least not in this sample.

Building and Interpreting

Create a scatter plot to determine whether there is a correlation between the percentage of fish eaten and the number of trips to the doctor in the past 6 months.

```
plot(x = sick$perc_fish, y = sick$doctor_trips)
```



How do we interpret this plot?