Module 2-1

Learning Objectives

1. Students will become familiar with the real-world problem for this module - food poisoning due to an outbreak of disease in one of the species of fish that is grown in recirculating aquaculture systems.

2. Students will be be able to generate two types of plots using base R syntax to visualize single variables (histograms and distributions) and two variables (scatterplots).
3. Students will develop visual-thinking skills to create visualizations that allow them to explore patterns in data, draw inferences and create solutions.

Introduction to the problem

Overall, the issue here is that 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, and it seems 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).

Diagram this next part on the board

We have a good set of data on team members that are sick, and how much fish or plant material they incorporate into their diets (team members vary in the composition of their diet - people are allowed to choose how much of different food sources they use).

Discussion - how might we figure this out?

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

Introduction to the dataset and visualization in R

First we're going to pull in the data and give it a quick inspection/exploration before we start to work through some of the visualization tools in R.

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

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

```
sick = read_csv("https://tinyurl.com/unavf2v")
```

```
## Parsed with column specification:
## cols(
##   last = col_character(),
##   first = col_character(),
##   gender = col_character(),
##   age = col_double(),
##   weight = col_double(),
##   specialties = col_character(),
##   perc_fish = col_double(),
##   perc_plant = col_double(),
##   doctor_trips = col_double()
## )
```

```
sick
```

```
## # A tibble: 362 x 9
##    last  first gender   age weight specialties perc_fish perc_plant doctor_trips
##    <chr> <chr> <chr>  <dbl>  <dbl> <chr>           <dbl>      <dbl>        <dbl>
##  1 al-R~ Faat~ M         22  194   Geology         0.572     0.428            3
##  2 Haas~ Hann~ F         55  236.  Electrical~     0.822     0.178            4
##  3 Asue~ Jasm~ F         64  211   Genetics        0.902     0.0984           5
##  4 Le    Harl~ F         38  114.  Applied Bi~     0.203     0.797            1
##  5 Milb~ Asien M         68  134   Data Scien~     0.586     0.414            3
##  6 Abey~ Char~ M         21  150   Mechanical~     0.294     0.706            1
##  7 Cimi~ Laur~ F         45  138.  Geology         0.372     0.628            2
##  8 Evins Ethan M         16  120.  Geology         0.192     0.808            2
##  9 Sanc~ Alon~ M         63  238.  Marine Bio~     0.353     0.647            1
## 10 Ange~ Andre M         39   92.5 Horticulure     0.172     0.828            1
## # ... with 352 more rows
```

```
glimpse(sick)
```

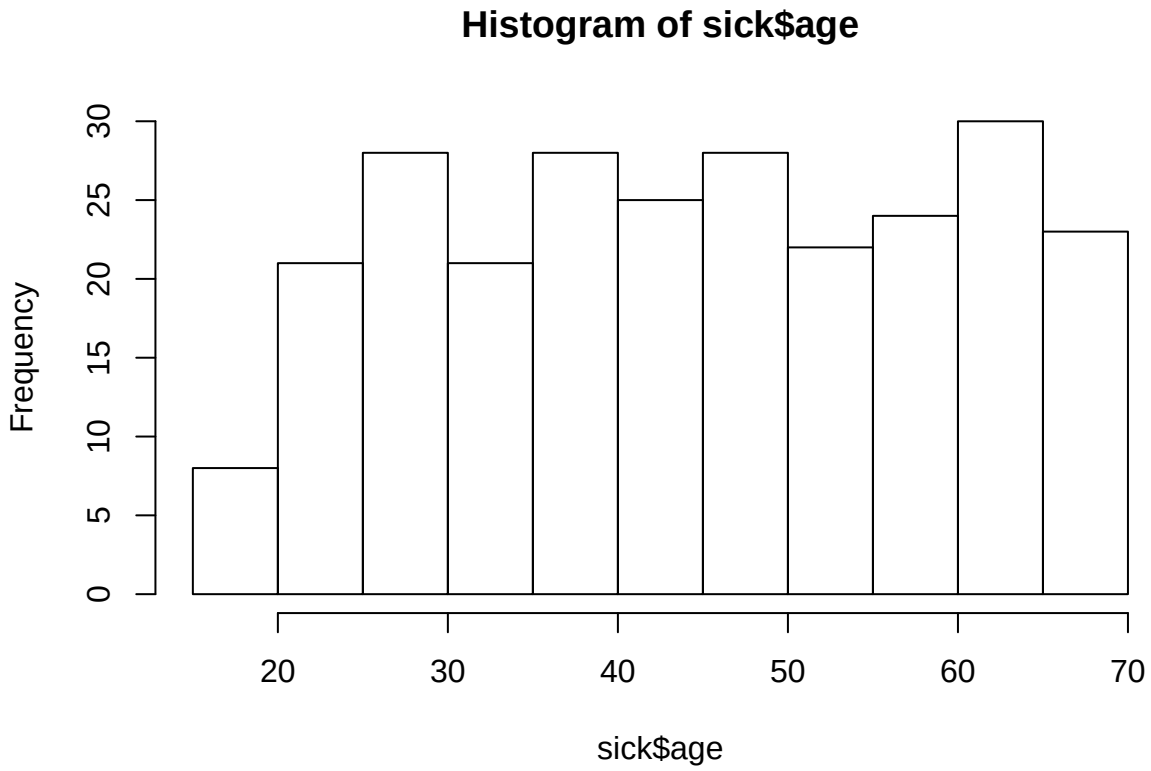
```
## Observations: 362
## Variables: 9
## $ last         <chr> "al-Rayes", "Haasenritter", "Asuega", "Le", "Milburn",...
## $ first        <chr> "Faatin", "Hannah", "Jasmine", "Harleigh", "Asien", "C...
## $ gender       <chr> "M", "F", "F", "F", "M", "M", "F", "M", "M", "M", "F",...
## $ age          <dbl> 22, 55, 64, 38, 68, 21, 45, 16, 63, 39, 50, 57, 63, 69...
## $ weight       <dbl> 194.0, 235.5, 211.0, 114.5, 134.0, 150.0, 138.5, 119.5...
## $ specialties  <chr> "Geology", "Electrical Engineering", "Genetics", "Appl...
## $ perc_fish    <dbl> 0.571604841, 0.822218933, 0.901635382, 0.202639454, 0....
## $ perc_plant   <dbl> 0.42839516, 0.17778107, 0.09836462, 0.79736055, 0.4143...
## $ doctor_trips <dbl> 3, 4, 5, 1, 3, 1, 2, 2, 1, 1, 2, 1, 2, 1, 3, 1, 1, 2, ...
```

```r
# See any problems? Looks like there might be some age data that is messed up
# some of the age values are wayyyy too young. Let's exclude all of those.

sick = sick %>%
  filter(age > 18)

# Basic visualization - in base R, the plot(), and some others function is your
# friend. We'll talk about some tidyverse ways to plot that are a bit more
# powerful later, but you really can do any kind of plotting you want with base
# R.

hist(sick$age)
```



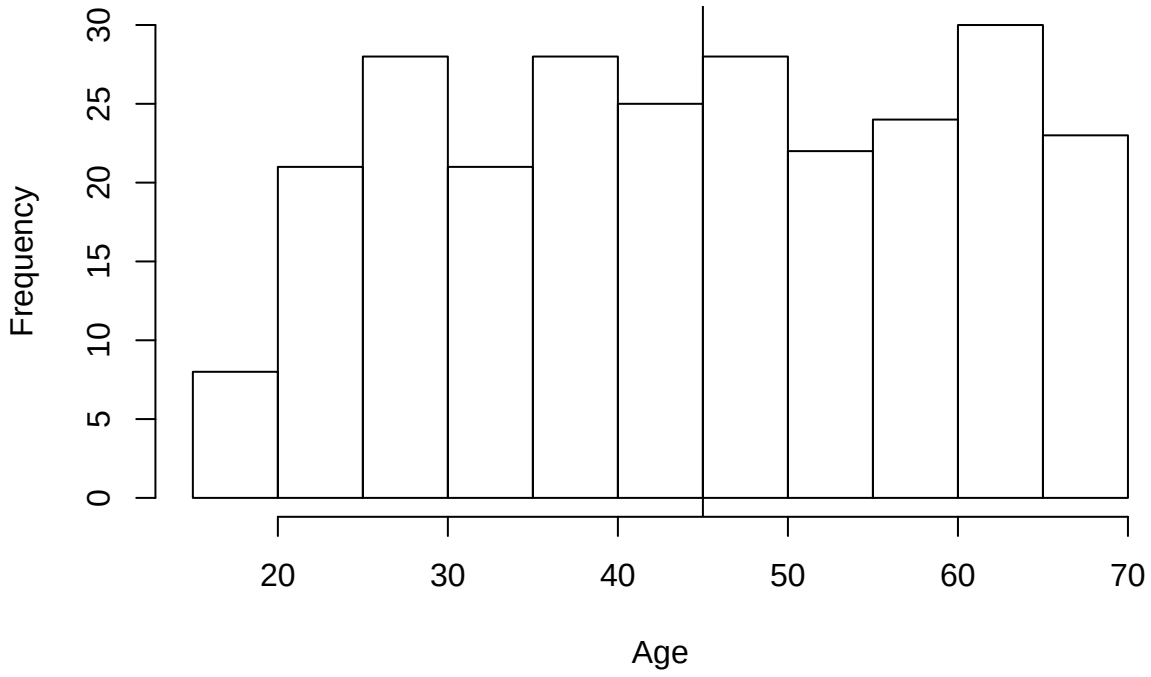
Mini Group Discussion - Histograms

What is this showing us? What conclusions can we draw from this data visualization? How could this visualization be improved?

Group Project - make a better histogram.

Work in your groups to build a histogram of weights with properly labeled axes, a vertical line that depicts the median, and an appropriate visually appealing bin size. Check out the `hist()` help page, and the `abline()` function.

```r
hist(sick$age, breaks = 15, xlab = "Age", ylab = "Frequency", main = NULL)
abline(v = median(sick$age))
```

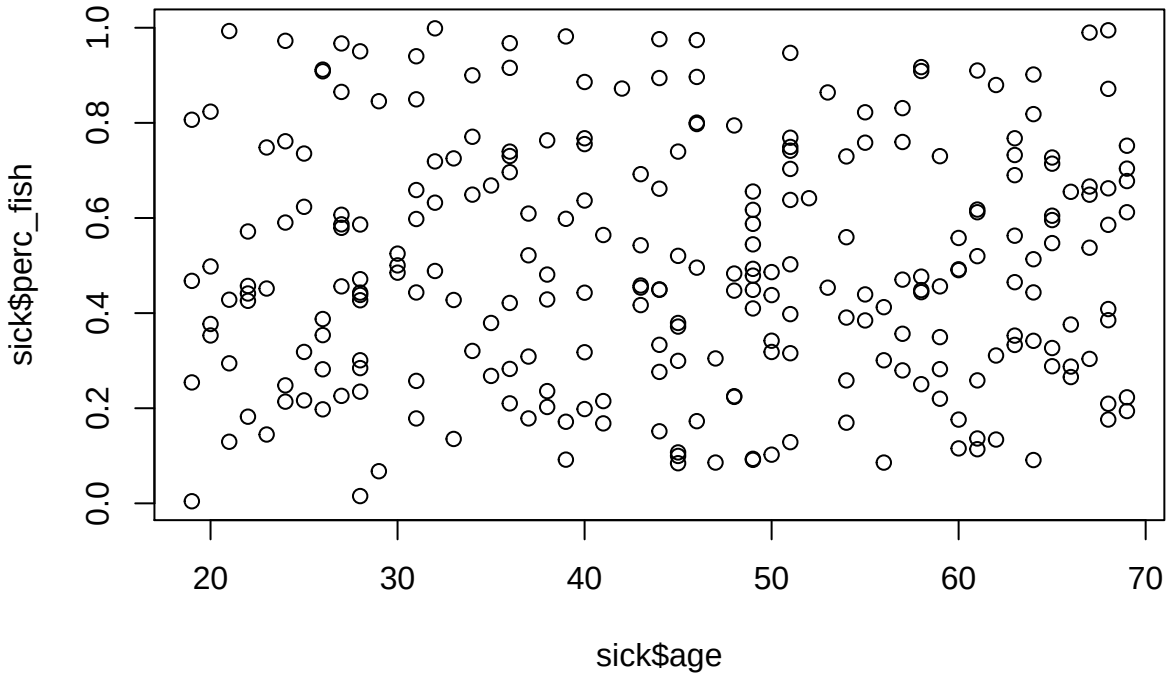


Brainstorm - Group activity

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 10 minutes in your groups sketching out a visualization that might give us insight into this.

Scatter plots

```r
plot(x = sick$age, y = sick$perc_fish)
```

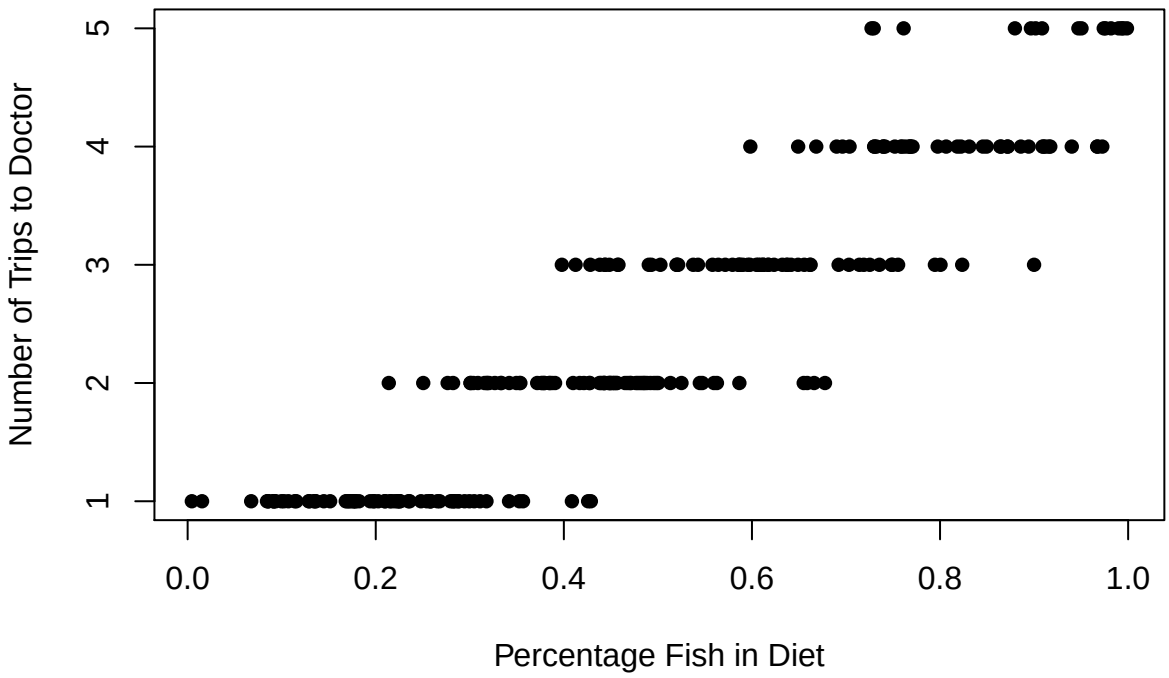


Not super informative, because there isn't really a relationship between a person's age and the amount of fish they consume. At least not in this sample.

Building and interpreting

In your group, build out the plot that you conceived earlier that determines whether there is a correlation between the percentage of fish eaten and the number of trips to the doctor in the past 6 months. Use proper labelling and aesthetic design principles to make it as visually appealing as possible.

```
plot(x = sick$perc_fish, y = sick$doctor_trips,
     xlab = "Percentage Fish in Diet",
     ylab = "Number of Trips to Doctor",
     main = NULL,
     pch = 16)
```


Number of Trips to Doctor
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
Percentage Fish in Diet