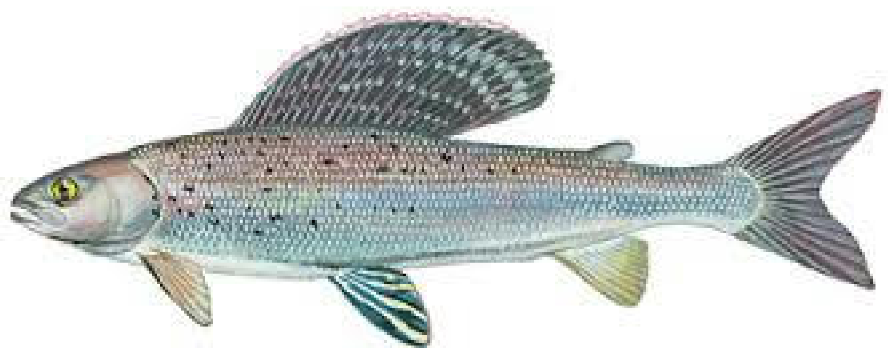
03_Class_Activity

Bill Perry

In class activity 3:



What did we do last time?

- Implement data pipeline best practices
- Apply controlled vocabulary and naming conventions
- Create effective visualizations
- Customize plots for publication quality
- Combine multiple plots into composite figures

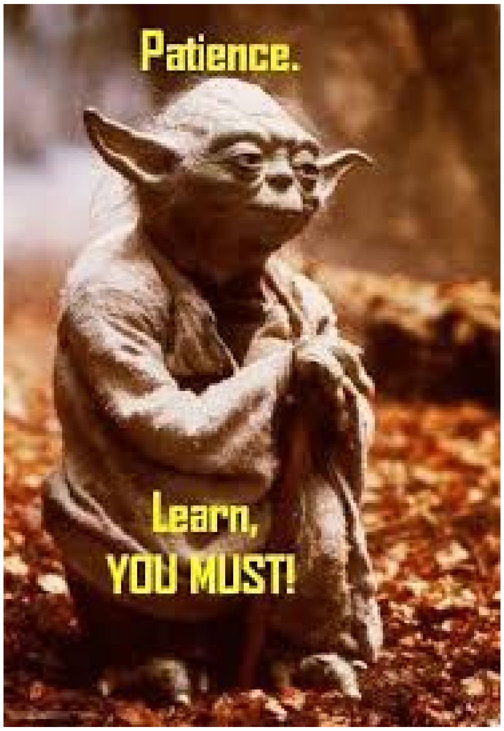
```
ggplot(name_df, aes(x_variable, y_variable, color = categorical_variable)) +
#      dataframe, aesthetics(x and y variables, mapping of color or fill or shape) +
  geom_point() +
# this it the geometry you want and can add more layers like
  geom_line()
```

- What questions do you have and what is unclear
- What did not work so far when you started the homework?

Objectives and goals for today

Today's Objectives

1. Implement descriptive statistics in R
2. Calculate measures of central tendency and spread
3. Compare distributions of data from different groups
4. Create effective visualizations of descriptive statistics
5. Interpret the meaning of these statistics in a biological context



Part 1: Setting Up Your Environment

First, let's load the necessary packages and import our data:

```r
# Load required packages

library(knitr)        # For creating tables
library(moments)      # For calculating skewness and kurtosis
library(skimr)        # for summary stats
library(flextable)    # for tables if you want - now tinytable
library(tidyverse)    # For data wrangling and visualization

# Set a consistent theme for our plots
theme_set(theme_minimal(base_size = 12))
```

Getting the data

Practice Exercise 1: Loading and Examining the Grayling Data

We'll be working with data on arctic grayling fish from two different lakes (I3 and I8).

```
# Write your code here to read in the file
# How do you examine the data - what are the ways you think and lets try it!

# Load the grayling data
g_df <- read_csv("data/gray_I3_I8.csv")
```

```
Rows: 168 Columns: 5
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (2): lake, species
dbl (3): site, length_mm, mass_g

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.
```

```
# View the first few rows
head(g_df)
```

```
# A tibble: 6 × 5
   site lake  species         length_mm mass_g
  <dbl> <chr> <chr>               <dbl>  <dbl>
1   113 I3    arctic grayling       266    135
2   113 I3    arctic grayling       290    185
3   113 I3    arctic grayling       262    145
4   113 I3    arctic grayling       275    160
5   113 I3    arctic grayling       240    105
6   113 I3    arctic grayling       265    145
```

```
# Examine the data structure
glimpse(g_df)
```

```
Rows: 168
Columns: 5
$ site      <dbl> 113, 113, 113, 113, 113, 113, 113, 113, 113, 113, 113, 113, …
$ lake      <chr> "I3", "I3", "I3", "I3", "I3", "I3", "I3", "I3", "I3", "I3", …
$ species   <chr> "arctic grayling", "arctic grayling", "arctic grayling", "ar…
$ length_mm <dbl> 266, 290, 262, 275, 240, 265, 265, 253, 246, 203, 289, 239, …
$ mass_g    <dbl> 135, 185, 145, 160, 105, 145, 150, 130, 130, 71, 179, 108, 1…
```

```
# Get a statistical summary
summary(g_df)
```

```
      site         lake              species            length_mm
 Min.   :113   Length:168         Length:168         Min.   :191.0
 1st Qu.:113   Class :character   Class :character   1st Qu.:270.8
```

```
 Median :118   Mode  :character   Mode  :character   Median :324.5
 Mean   :116                                         Mean   :324.5
 3rd Qu.:118                                         3rd Qu.:377.0
 Max.   :118                                         Max.   :440.0

     mass_g
 Min.   : 53.0
 1st Qu.:151.2
 Median :340.0
 Mean   :351.2
 3rd Qu.:519.5
 Max.   :889.0
 NA's   :2
```

```r
# How many fish do we have from each lake?

g_df %>%
  count(lake)
```

```
# A tibble: 2 × 2
  lake      n
  <chr> <int>
1 I3       66
2 I8      102
```

Questions to Consider:

1. What variables are in our dataset?
2. What are their data types?
3. Are there any missing values?
4. What is the range of fish lengths in our dataset?
5. How many fish were sampled from each lake?

Part 2: Calculating Descriptive Statistics

Let's calculate various descriptive statistics for our data:

Practice Exercise 2: Measures of Central Tendency

Let's recreate the basic histogram of fish lengths from our last class. Use the `sculpin_df` data frame that's already loaded.

```
# Write your code here to read in the file
# How do you examine the data - what are the ways you think and lets try it!
# Calculate the mean and median fish length
mean(g_df$length_mm)
```

```
[1] 324.494
```

```
median(g_df$length_mm)
```

```
[1] 324.5
```

```
# Calculate mean and median by lake
g_df %>%
  group_by(lake) %>%
  summarise(
    mean_length = mean(length_mm),
    median_length = median(length_mm)
  )
```

```
# A tibble: 2 × 3
  lake  mean_length median_length
  <chr>       <dbl>         <dbl>
1 I3           266.           266
2 I8           363.           373
```

Summarizing data - two ways

lets say we want to summarize the data and need to get n, means, standard deviation, standard error

We could do the following - if we had missing cells the code below would give an error

```
mean(g_df$length_mm)
```

```
[1] 324.494
```

```
mean(g_df$length_mm, na.rm = TRUE) # removes missing values
```

```
[1] 324.494
```

```
length(g_df$length_mm)
```

```
[1] 168
```

- **the length counts missing and non-missing data**
- however this would get old if we had to do this for everything and then to do it for the different groupings - lee and windward...

We need to learn to pipe

passes things from the dataframe to a command and so on...

- the dataframe –> pipe command that feed the dataframe into –> next command

```
g_df %>% summarize(mean_length = mean(length_mm, na.rm = TRUE))
```

```
# A tibble: 1 × 1
  mean_length
        <dbl>
1        324.
```

What is cool is we can do a lot of different things now

```
g_df %>%
  summarize(
    mean_length = mean(length_mm, na.rm = TRUE),
    sd_length = sd(length_mm, na.rm = TRUE),
    n_length = n())
```

```
# A tibble: 1 × 3
  mean_length sd_length n_length
        <dbl>     <dbl>    <int>
1        324.      65.0      168
```

Super cool code in case there are missing values

```
g_df %>%
  summarize(
    mean_length = mean(length_mm, na.rm = TRUE),
    sd_length = sd(length_mm, na.rm = TRUE),
    n_length = sum(!is.na(length_mm)))
```

```
# A tibble: 1 × 3
  mean_length sd_length n_length
        <dbl>     <dbl>    <int>
1        324.      65.0      168
```

Now for Spread...

Practice Exercise 3: Measures of Spread

```r
# Write your code here to read in the file
# Calculate standard deviation and variance
mean_length <- mean(g_df$length_mm, na.rm=TRUE)
sd_length <- sd(g_df$length_mm)
var_length <- var(g_df$length_mm)
sd_length
```

```
[1] 65.00659
```

```r
var_length
```

```
[1] 4225.856
```

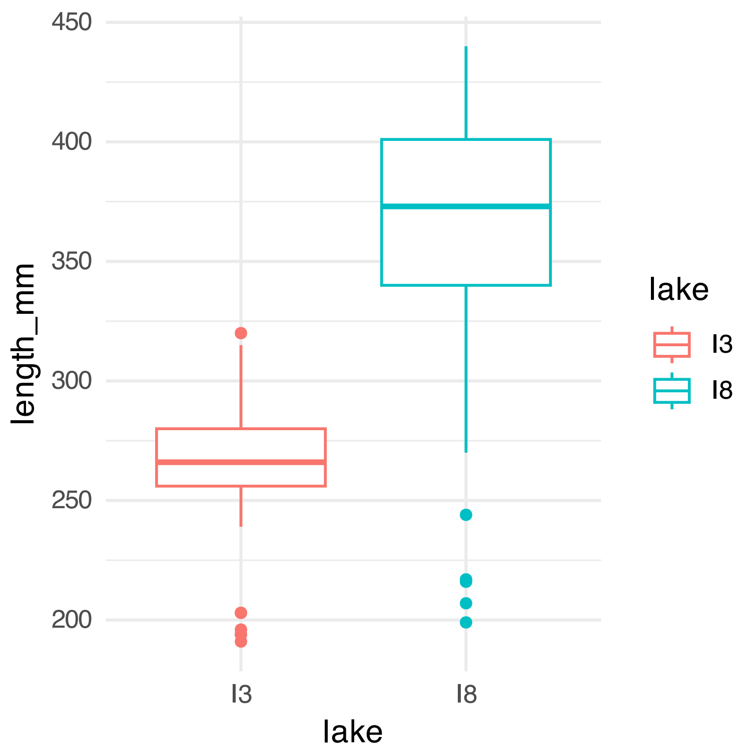
Exercise 4: Calculate Quartiles and Percentiles

```r
# Calculate quartiles for overall data
quartiles <- quantile(g_df$length_mm, probs = c(0.25, 0.5, 0.75))
# cat("First quartile (Q1):", quartiles[1], "mm\n")
# cat("Second quartile (Median):", quartiles[2], "mm\n")
# cat("Third quartile (Q3):", quartiles[3], "mm\n")

# Calculate a more comprehensive set of percentiles
percentiles <- quantile(g_df$length_mm,
                        probs = c(0.1, 0.25, 0.5, 0.75, 0.9))

# Display the percentiles using flextable
data.frame(
  Percentile = c("10th", "25th (Q1)", "50th (Median)", "75th (Q3)", "90th"),
  Value = percentiles
)
```

```
       Percentile  Value
10%          10th 251.10
25%     25th (Q1) 270.75
50% 50th (Median) 324.50
75%     75th (Q3) 377.00
90%          90th 408.60
```

Note you could add a box plot by lake to see this if you wanted

```r
g_df %>%
  ggplot(aes(lake, length_mm, color= lake))+
  geom_boxplot()
```



Exercise 5: Calculate the Coefficient of Variation

The coefficient of variation (CV) is the standard deviation expressed as a percentage of the mean:

$$CV = \frac{s}{\bar{Y}} \times 100\%$$

```
# Calculate coefficient of variation
sd_length / mean_length * 100
```

```
[1] 20.03321
```

```
# Calculate by lake
g_df %>%
  group_by(lake) %>%
  summarise(
    mean_length = mean(length_mm),
    sd_length = sd(length_mm),
    cv_length = sd_length / mean_length * 100
  ) %>%
  flextable()
```

lake	mean_length	sd_length	cv_length
I3	265.6061	28.30378	10.65630
I8	362.5980	52.33901	14.43444

Questions to Consider:

1. How do the means and medians compare within each lake? What might this tell you about the distribution?

2. Which lake has more variable fish lengths? How can you tell?
3. Why might the coefficient of variation be useful when comparing variability between different measurements (e.g., length vs. mass)?

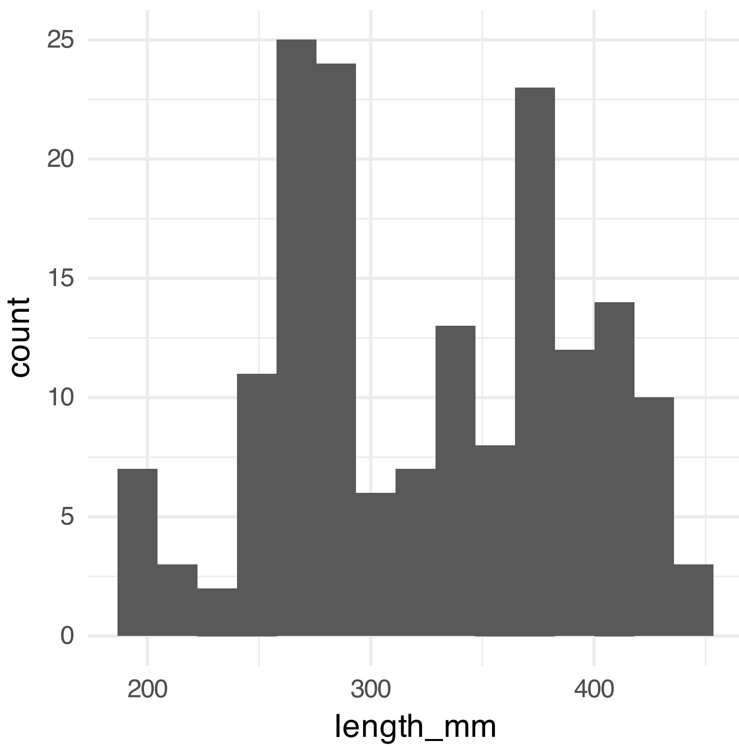
Part 3: Visualizing Distributions

Visualizations can help us better understand the descriptive statistics we've calculated.

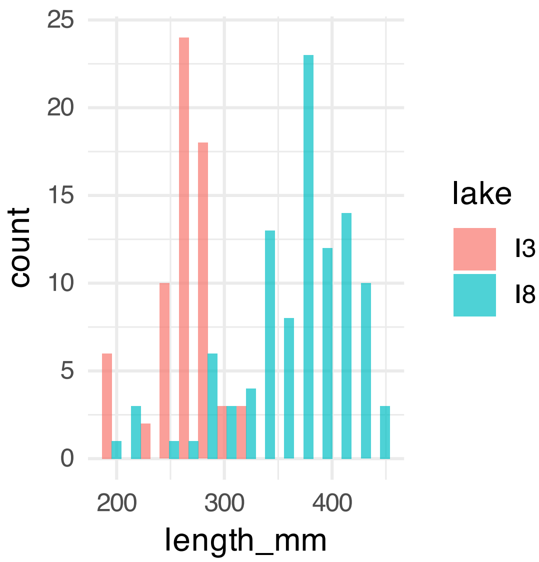
Exercise 6: Creating Histograms

One of the best ways to look at data is a histogram - and we will do it again

```
# Create a histogram of all fish lengths
g_df %>% ggplot(aes(x = length_mm)) +
  geom_histogram(bins = 15)
```



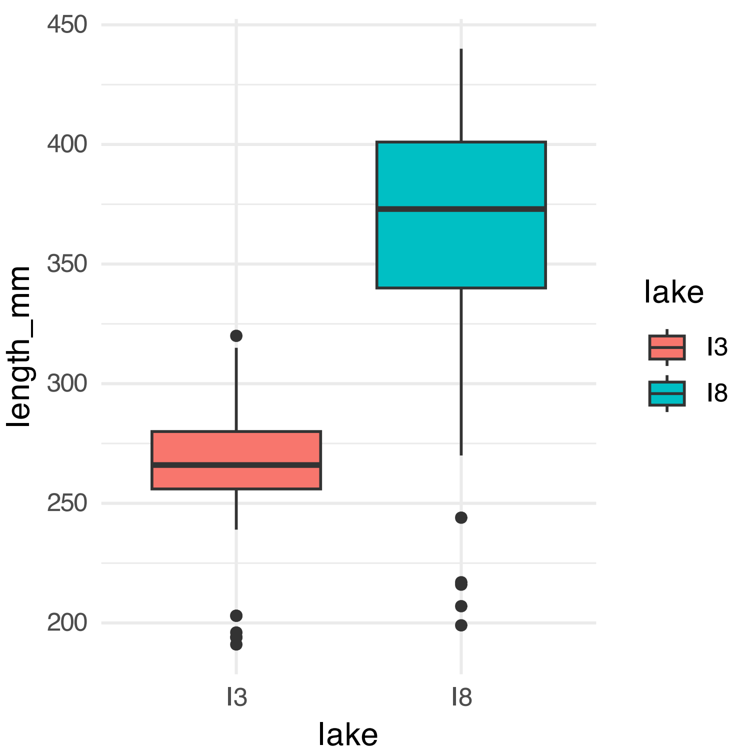

```
# Create histograms by lake
g_df  %>% ggplot(aes(x = length_mm, fill = lake)) +
  geom_histogram(bins = 15, position = "dodge", alpha = 0.7)
```



Exercise 7: Creating Box Plots

Personally I like box plots

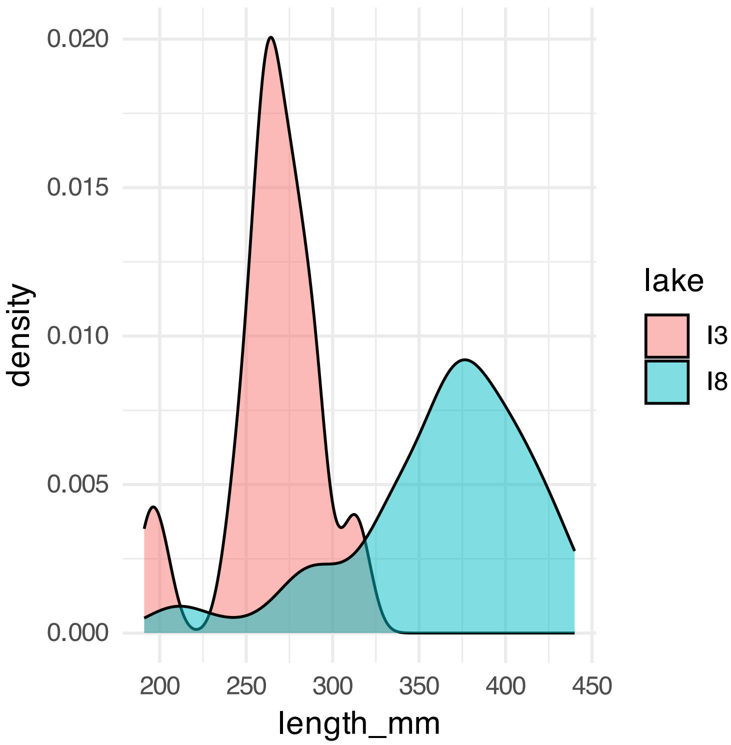
```
# Create a box plot comparing fish lengths by lake
# Create a box plot comparing fish lengths by lake
g_df  %>%  ggplot( aes(x = lake, y = length_mm, fill = lake)) +
  geom_boxplot()
```



Exercise 9: Creating Density Plots

Now these will be really important later on

```
## Create density plots
g_df  %>%  ggplot(aes(x = length_mm, fill = lake)) +
  geom_density(alpha = 0.5)
```



Questions to Consider:

1. Which visualization best shows the differences in fish lengths between lakes?
2. What can you learn from the violin plots that might not be apparent from the box plots?
3. How would you interpret the cumulative frequency distribution?
4. What patterns or insights can you identify from these visualizations?

Part 4: Interpreting the Results

Based on our analysis, we can make the following observations:

1. **Lake Differences**: Fish from Lake I8 are generally larger than those from Lake I3, both in length and mass.
2. **Variability**: Lake I8 shows greater variability in fish lengths and masses than Lake I3, as indicated by higher standard deviations and IQRs.
3. **Distribution Shape**:
   - Lake I3 fish lengths are more symmetrically distributed.
   - Lake I8 fish lengths show a slight negative skew, suggesting a few smaller fish pulling the distribution to the left.
4. **Length-Mass Relationship**: Both lakes show a strong positive correlation between fish length and mass, following an approximately cubic relationship (mass increases with the cube of length).

Guided Questions for Deeper Understanding of descriptive statistics

1. **Biological Interpretation**: What ecological factors might explain the differences in fish size between the two lakes?
2. **Statistical Reasoning**: Why might we prefer to use the median and IQR instead of the mean and standard deviation in some cases?
3. **Data Visualization**: Which visualization method was most effective for comparing the two lakes? Why?
4. **Scientific Communication**: How would you concisely summarize these findings in a scientific paper?
5. **Further Analysis**: What additional analyses might be useful to better understand this dataset?