PSTAT 5LS Lab 1

TA Name Here

Summer 2025

Section 1

Learning Objectives

R Learning Objectives

1. Learn how to import data into R
2. Learn how to find the five-number summary of a variable, and find a specific numeric summary (statistic) in R
3. Learn how to make a histogram in R
4. Learn how to make a box plot in R
5. Learn how to make side-by-side box plots in R

Statistical Learning Objectives

1. Understand when to histogram
2. Understand when to make a box plot
3. Understand when to make a side-by-side box plot and how to use this type of comparison
4. Be able to use these graphical and numerical summaries to discuss data

Functions covered in this lab

1. `read.csv()`
2. `head()`
3. `str()`
4. `summary()`
5. `hist()`
6. `min()`, `mean()`, `median()`, `max()`, `sd()`, `IQR()`
7. `boxplot()`

Section 2

Lab Tutorial

How Data Can Be Stored: CSV Files

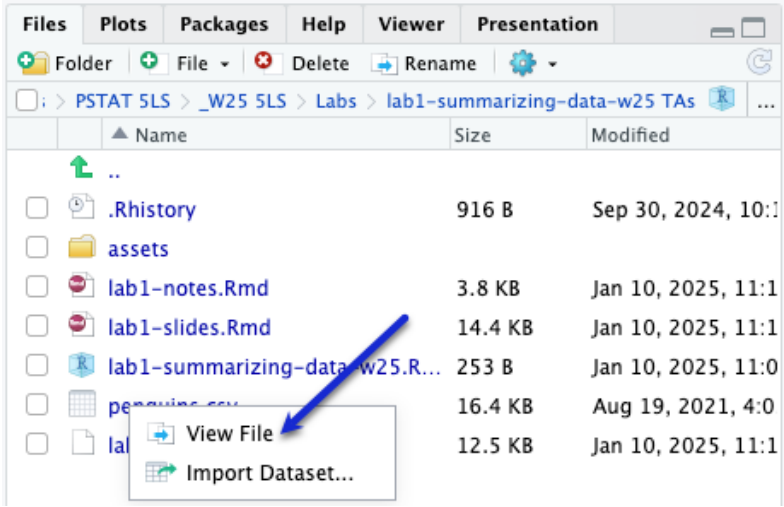
One common way to store data is to store it in a **CSV file**. CSV stands for "Comma Separated Values".

Structure of a CSV File

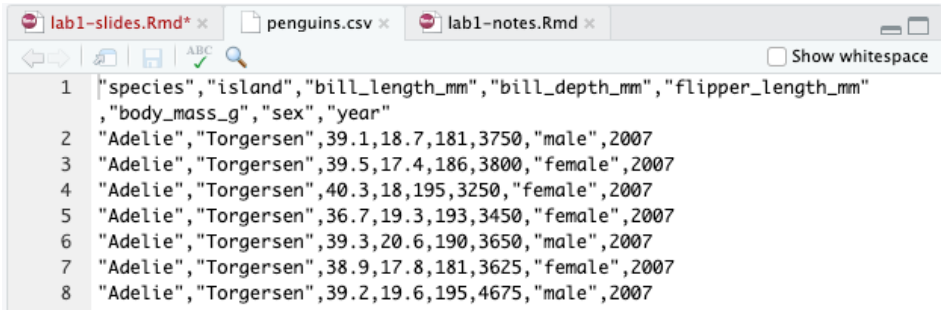
- **Header Row:** The first row lists the names of the variables in the file.
- **Subsequent Rows:** Each row of the file is an "observation" or "case", and consists of one or more variables whose *values* are *separated* by *commas*.

Explore a Real CSV File!

- Open the file **"penguins.csv"** from the Files pane (lower right).



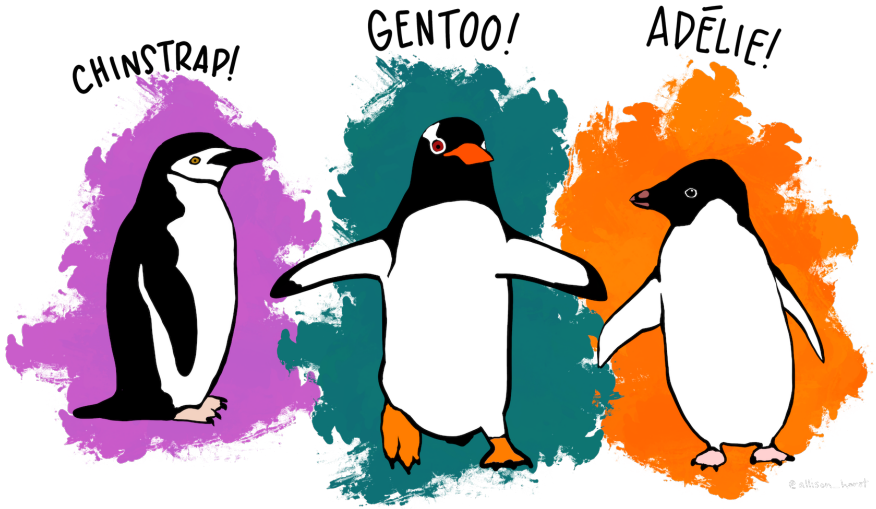
- Observe how the rows and columns are organized.



- CSV files are a simple yet powerful way to organize and share data!

Palmer Penguins Data

We'll work with a data set of 333 penguins collected from 3 islands in the Palmer Archipelago in Antarctica. Ther data, collected by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER, was prepared by Dr. Allison Horst.



Reading Data into R

Using `read.csv()`

- We use the function `read.csv()` to **read data** into R.
- The first argument is the name of a `.csv` file (in quotes), e.g., `"penguins.csv"`.
- The results of `read.csv()` are stored in an object, here named `penguins`.

```
penguins <- read.csv("penguins.csv", stringsAsFactors = TRUE)
```

What Does `stringsAsFactors = TRUE` Do?

- **Strings:** Words or phrases in the data
- **Factors:** Levels of a categorical variable
- Setting `stringsAsFactors = TRUE` tells R to treat words or phrases as **categorical variables.**

Steps to Read in the Penguins Data

1. Run the `loadPenguins` chunk of your `lab1-notes.Rmd` file.
2. Check that the `penguins` data appears in your RStudio Environment (top right corner).

Tip: Always verify your data after loading it to ensure it's been imported correctly!

Peeking at the Data

We can peek at the first few (6, specifically) rows of the data using the `head()` function:

```
head(penguins)
```

```
##   species    island bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
## 1  Adelie Torgersen           39.1          18.7               181        3750
## 2  Adelie Torgersen           39.5          17.4               186        3800
## 3  Adelie Torgersen           40.3          18.0               195        3250
## 4  Adelie Torgersen           36.7          19.3               193        3450
## 5  Adelie Torgersen           39.3          20.6               190        3650
## 6  Adelie Torgersen           38.9          17.8               181        3625
##      sex year
## 1   male 2007
## 2 female 2007
## 3 female 2007
## 4 female 2007
## 5   male 2007
## 6 female 2007
```

The `penguins` data set contains a number of *variables* (e.g., `species`, `island`).

Use the `tryIt1` chunk in your notes to peek at the first 6 rows of the penguins data file. The function is `head(penguins)`.

Data

Variable name	Description
`species`	Penguin species (Adélie, Chinstrap, Gentoo)
`island`	Island in the Palmer Archipelago (Biscoe, Dream, Torgersen)
`bill_length_mm`	Bill length (in mm)
`bill_depth_mm`	Bill depth (in mm)
`flipper_length_mm`	Flipper length (in mm)
`body_mass_g`	Penguin body mass (in grams)
`sex`	Penguin sex (female, male)
`year`	Study year (2007, 2008, 2009)

Another Way to Peek at the Data

We can also peek at the data using the `str()` function (pronounced "stir", short for "structure").

`str()` shows the **structure** of the data set, including the types of variables and a preview of the data

```
str(penguins)
```

```
## 'data.frame':    333 obs. of  8 variables:
##  $ species          : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ island           : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ bill_length_mm   : num  39.1 39.5 40.3 36.7 39.3 38.9 39.2 41.1 38.6 34.6 ...
##  $ bill_depth_mm    : num  18.7 17.4 18 19.3 20.6 17.8 19.6 17.6 21.2 21.1 ...
##  $ flipper_length_mm: int  181 186 195 193 190 181 195 182 191 198 ...
##  $ body_mass_g      : int  3750 3800 3250 3450 3650 3625 4675 3200 3800 4400 ...
##  $ sex              : Factor w/ 2 levels "female","male": 2 1 1 1 2 1 2 1 2 2 ...
##  $ year             : int  2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
```

Use the `tryIt2` chunk in your notes examine the structure of the penguins data file.

How to Find Help in R

R has built-in documentation for every function. Instead of Googling, use R's help system: type `?function_name` in the console (e.g., `?hist`) to view the documentation.

In the `tryit3` chunk, try this out for the `hist()` function.

The help file often includes examples that you can run directly with `example(function_name)` (e.g., `example(hist)`).

The most useful part of the help file is the list of arguments and their descriptions. You may not understand everything right away, but give it a try and ask your TA if needed!

Summarizing the flipper_length_mm Variable

Let's start by looking at the flipper_length_mm variable. Is it categorical or quantitative? How can you tell?

We can summarize the data numerically using R. The summary() function provides a quick summary of any variable.

Let's summarize the flipper_length_mm variable, which represents the length of the penguins' flippers (in millimeters).

```
summary(penguins$flipper_length_mm)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     172     190     197     201     213     231
```

Try running this code in the tryit4 chunk in your notes to see the summary.

Numerical Summaries in R

The `summary()` function provides basic statistics, but it doesn't include the standard deviation. To get the standard deviation, use the `sd()` function.

Summarize the `flipper_length_mm` variable and include the standard deviation using the following code. Run the `tryit5` chunk in your notes.

```
summary(penguins$flipper_length_mm)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     172     190     197     201     213     231
```

```
sd(penguins$flipper_length_mm)
```

```
## [1] 14.01577
```

Specific Numerical Summaries

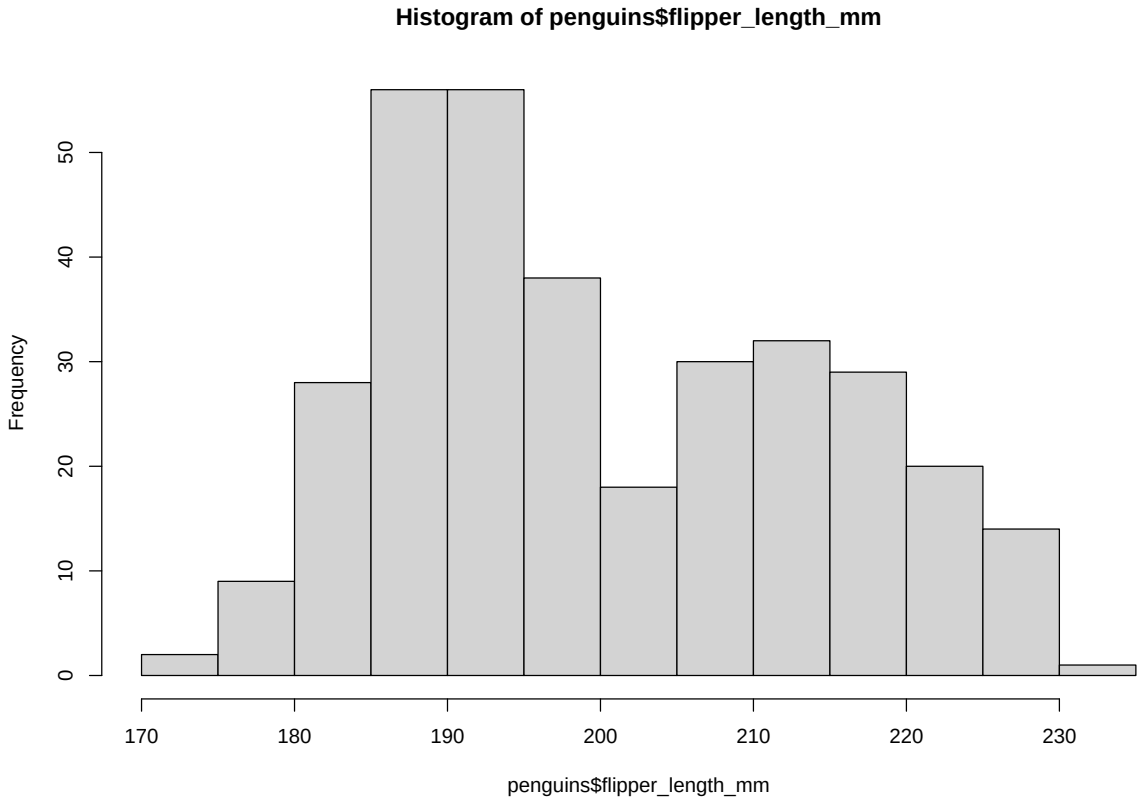
You can also get individual summary statistics using specific functions. Try the following in the `tryit6` code chunk.

```
min(penguins$flipper_length_mm)
mean(penguins$flipper_length_mm)
median(penguins$flipper_length_mm)
max(penguins$flipper_length_mm)
sd(penguins$flipper_length_mm)
IQR(penguins$flipper_length_mm)
```

Histograms in R

Histograms are used to visualize the distribution of a quantitative variable. You can easily create a histogram in R with the `hist()` function.

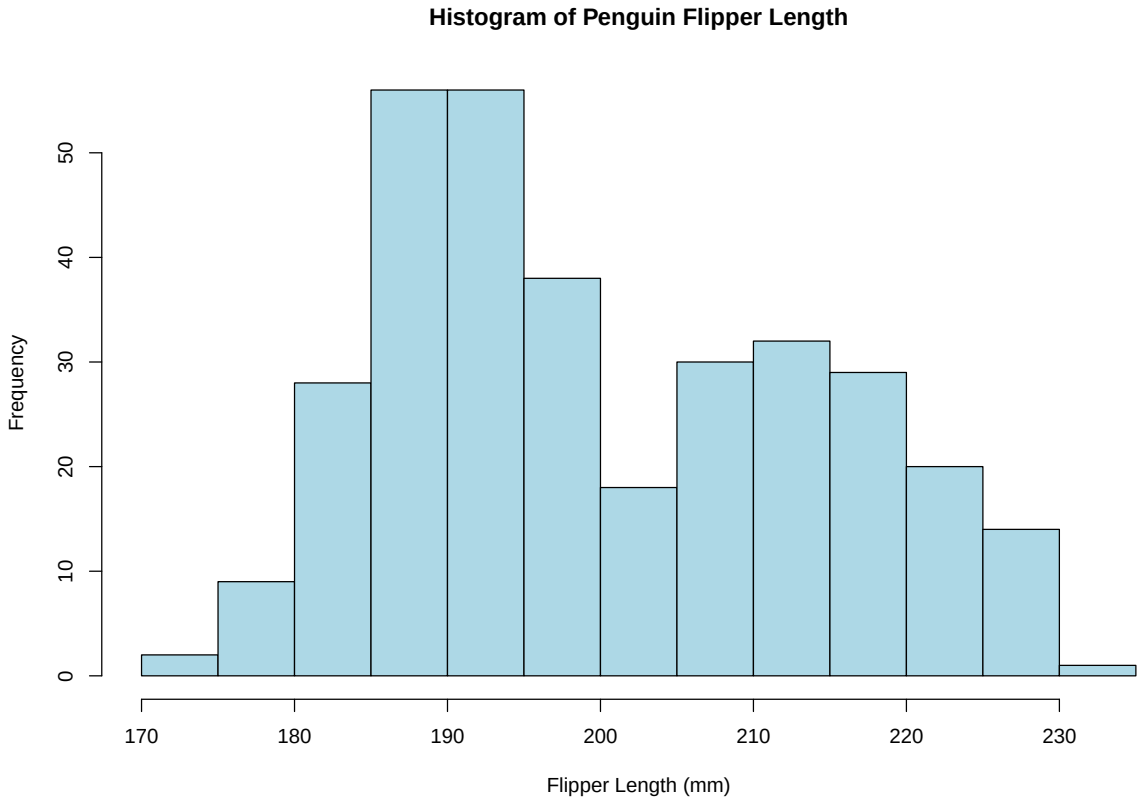
Example: Histogram of Flipper Length



Adding Labels and Titles

By default, R adds titles and axis labels, but they aren't always informative. Always include `main`, `xlab`, and `ylab` arguments to clarify your plot.

Here's an example of a better histogram with labels:



Try It Out!

Mark up and then run the code in the `tryit7` code chunk. Double-check for typos, as they can cause errors! If you encounter an error message, try to debug it yourself before asking for help.

```
hist(penguins$flipper_length_mm,
     main = "Histogram of Penguin Flipper Length",
     xlab = "Flipper Length (mm)",
     col = "lightblue")
```

Describing Histograms

As you learned in lecture, when describing a distribution, consider four key aspects:

1. Shape (modes and symmetry)
2. Center
3. Spread/Variability
4. Outliers

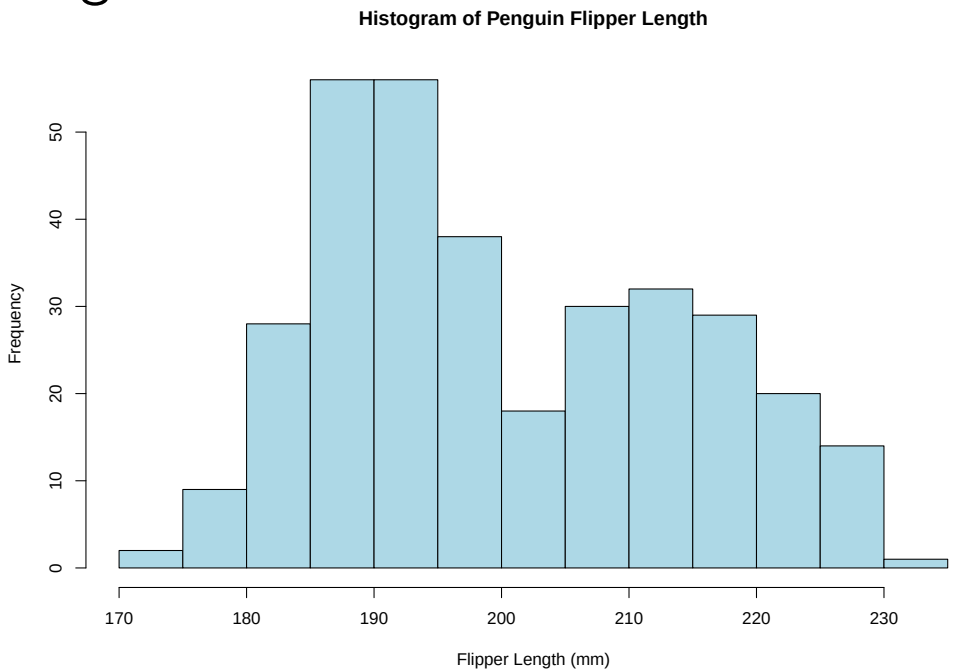
Use the mnemonic **SOCS**:

Shape **O**utliers **C**enter **S**pread

Note: Always mention whether outliers are present. Not addressing them suggests you missed checking for them.

Using Histograms to Describe Distributions

Here again is our histogram of flipper lengths:



Describe the distribution of flipper lengths.

Using Histograms to Describe Distributions

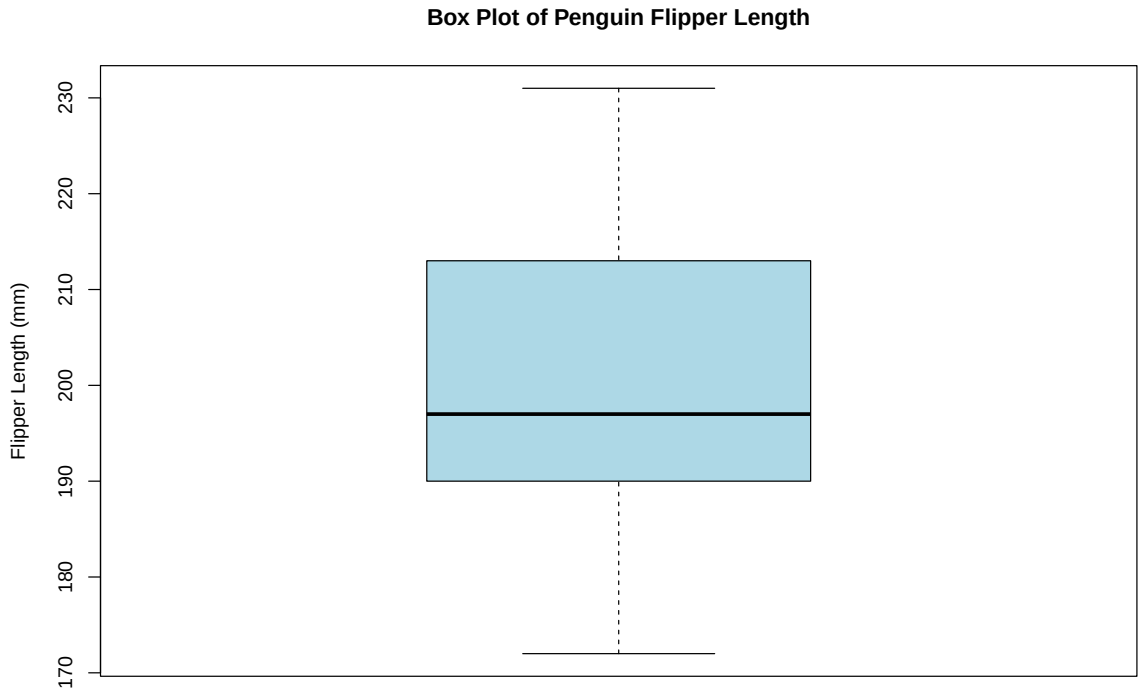
Do you think that the mean is the best measure of center for the flipper lengths? Why or why not?

Box Plots in R

A box plot is another effective way to visualize a quantitative variable.

Creating a box plot in R is straightforward: use the `boxplot()` function.

Just like with histograms, always include a title (`main`) and axis labels (`ylab`) to make your plot clear and informative.



Try It Out!

Mark up and then run the code in the `tryit8` code chunk. As before, watch out for typos. If you get an error message, try to debug it yourself before asking for help!

```
boxplot(penguins$flipper_length_mm,
        main = "Box Plot of Penguin Flipper Length",
        ylab = "Flipper Length (mm)",
        col = "lightblue")
```

Describing Distributions with Box Plots

True or False:

The box plot of flipper lengths appears to be unimodal and symmetric.

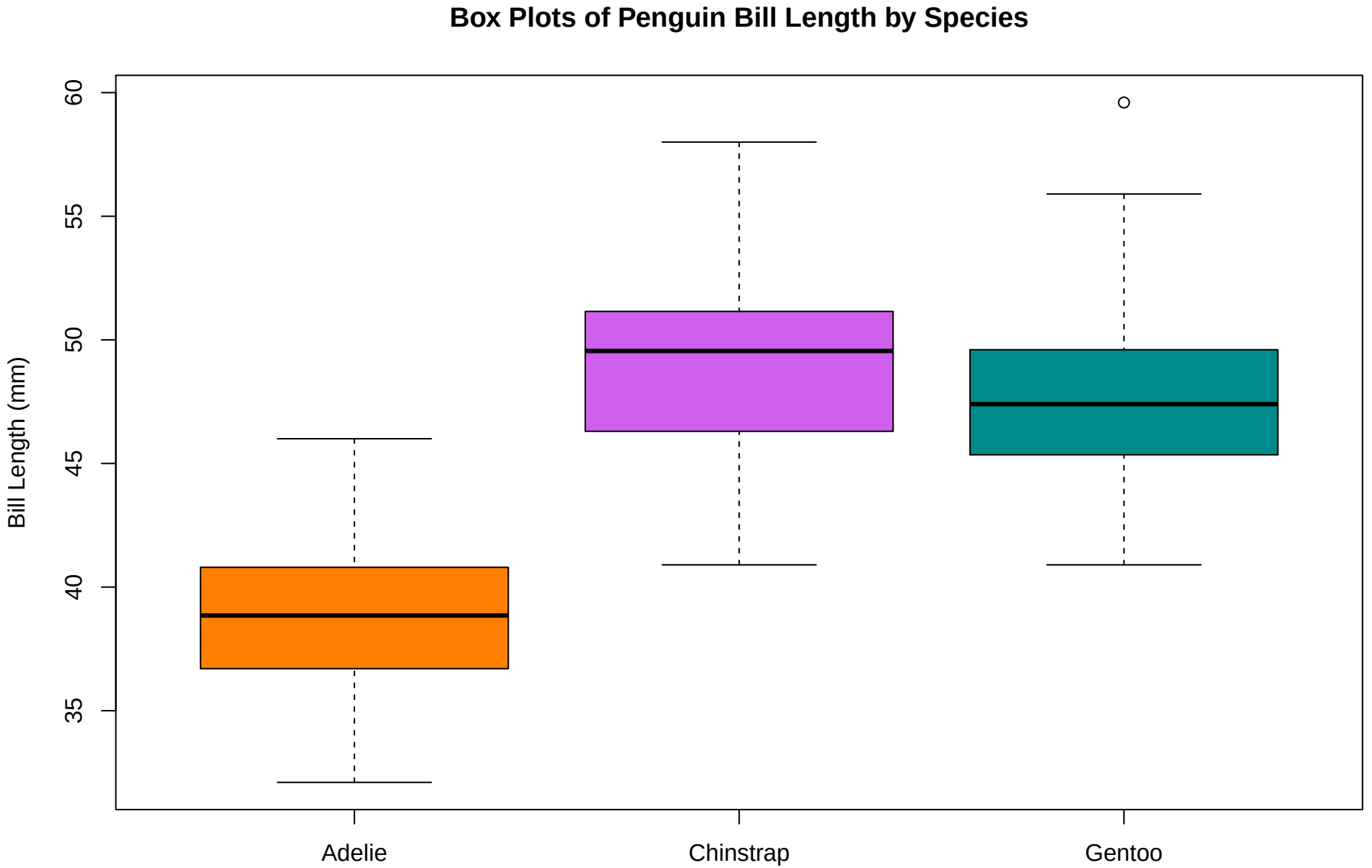
Side-by-side Box Plots

To compare groups, we we can use side-by-side box plots. For example, we can compare bill lengths across penguin species using the `boxplot()` function.

```
boxplot(penguins$bill_length_mm ~ penguins$species,
        main = "Box Plots of Penguin Bill Length by Species",
        ylab = "Bill Length (mm)",
        xlab = "Species",
        col = c("darkorange1", "mediumorchid2", "darkcyan"))
```

Try out the provided code in the `tryit9` chunk to generate these plots!

Side-by-side Box Plots Continued



Penguin Bill Length By Species

Does it appear that a penguin's bill length is related to its species, for the penguins in Palmer Archipelago? Why or why not?

What Next?

In today's lab, we used R to get graphical and numerical summaries for quantitative variables.

As we go throughout the quarter, we will continue learning how to analyze data.