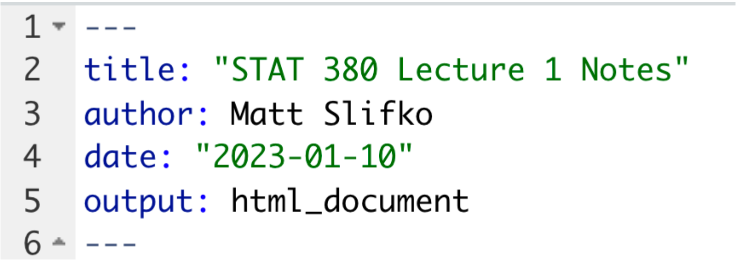
STAT 380 – Data Visualization Review (Lecture 1)

These notes largely follow from Chapters 1-3 of *Modern Data Science with R (MDSR)*.

R Markdown

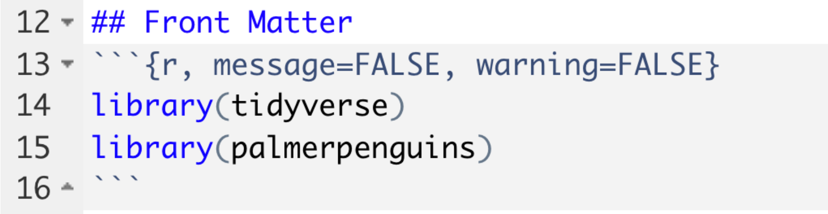
EXAMPLE 1: Most assignments in this class will be completed using R Markdown. Some expectations:

- Know how to start a new markdown document. To do so, go to RStudio >> File >> New File >> R Markdown and select OK.
- Know how to modify the YAML header in order to add your name and date. After adding "author" and "date" lines, a typical YAML header may appear as below:



```
---
title: "STAT 380 Lecture 1 Notes"
author: Matt Slifko
date: "2023-01-10"
output: html_document
---
```

- The document may have some default headers, text, and R chunks. Delete these.
- Add a Level 2 Header titled "Front Matter". In this section, add an R chunk and include all library commands and commands to read in datasets (e.g., read.csv, read.table, etc.) in this section.



~~~
## Front Matter
```{r, message=FALSE, warning=FALSE}
library(tidyverse)
library(palmerpenguins)
```
~~~

- Avoid displaying complete datasets in your markdown reports. While it is common to use the View() function in the console, you should NOT use the View() function in an R chunk. Similarly, avoid including the name of a dataset for the sake of viewing it in your R chunk (see example in STAT380_L1.Rmd file). Instead, use head(), glimpse(), str(), etc. in R chunks.
- In STAT 184, you may have used R Notebooks instead of R Markdown. While they are similar, there are a few minor differences. You are welcome to use either in this class.
    - If using an R Notebook, be sure that you remember to run all R chunks; otherwise plots, results, etc. may not appear in the html document.

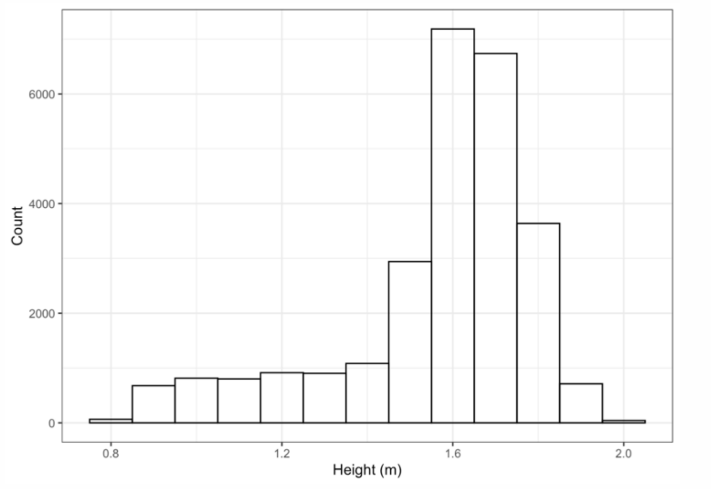
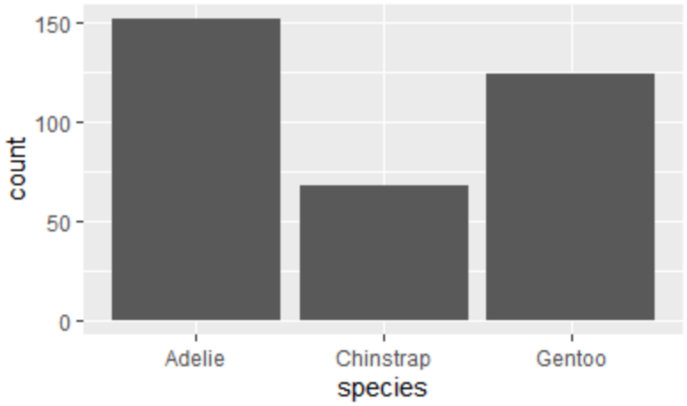
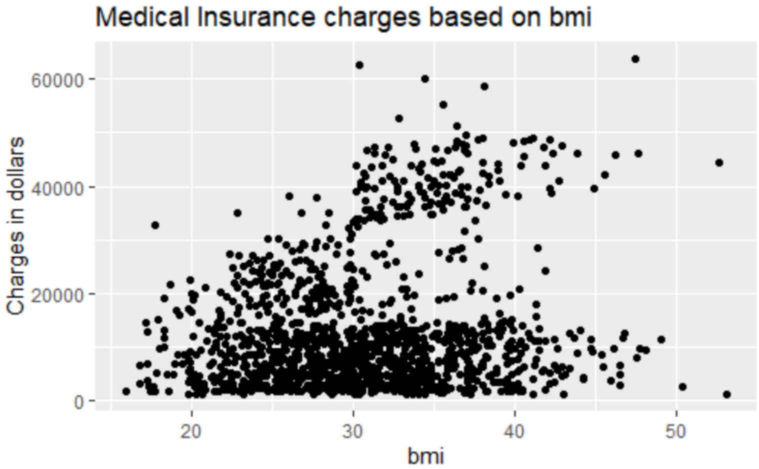
Data Visualization

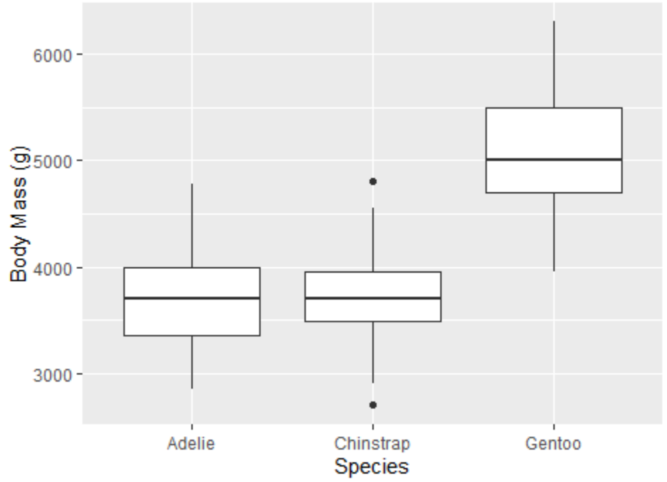
IDEA: Data graphics provide one of the most accessible, compelling, and expressive modes to investigate and depict patterns in data.

IDEA: Throughout this course, we will learn a variety of tools for various purposes (data wrangling, visualizations, modeling, validation, etc.). An important recurring theme is that you must be able to choose the appropriate tool. This decision is often based on the type of data that you are using.

RECALL: While you may have started creating visualizations using mplot or esquisse, functionality is limited. We will focus on writing our own ggplot plot code. ggplot is one of the most commonly used data visualization tools in R. Specifically, ggplot() is a function within the ggplot2 package.

EXAMPLE 2: A few common data visualizations produced using ggplot are shown below:

Plot	Name	Use When
	Histogram – shows how many cases fall into a given range of the variable	
	Bar Chart/Plot – shows how many cases (or relative proportions) of each category	
	Scatterplot – shows the relationship between two variables	

 	Side-by-side boxplots – depicts the five-number summary (min, Q1, median, Q3, and max) for each level of a categorical variable	

NOTE: Although there are many types of data visualizations, there are several canonical data graphics that are simplistic, yet very effective. Table 3.3 from MDSR, reproduced below for your convenience, provides a nice summary of both univariate and multivariate visualizations.

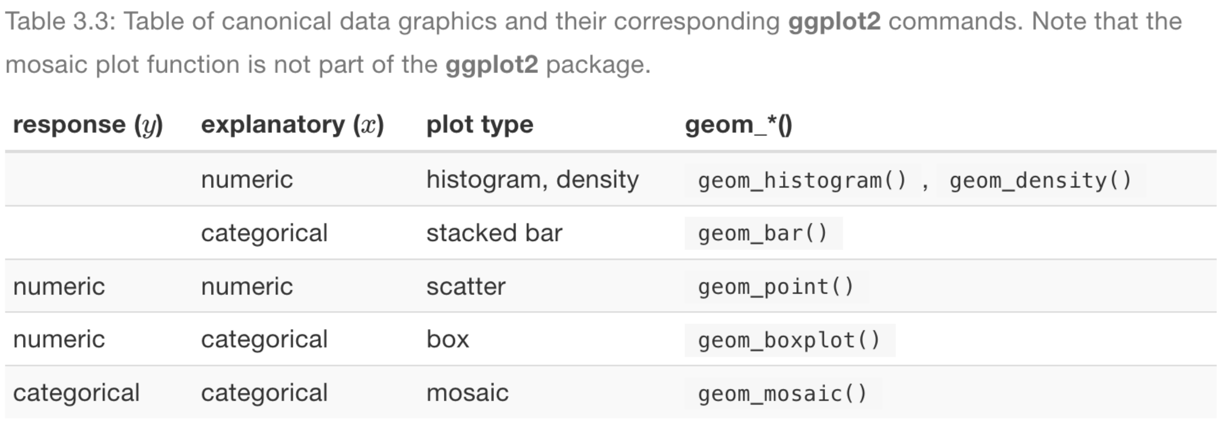
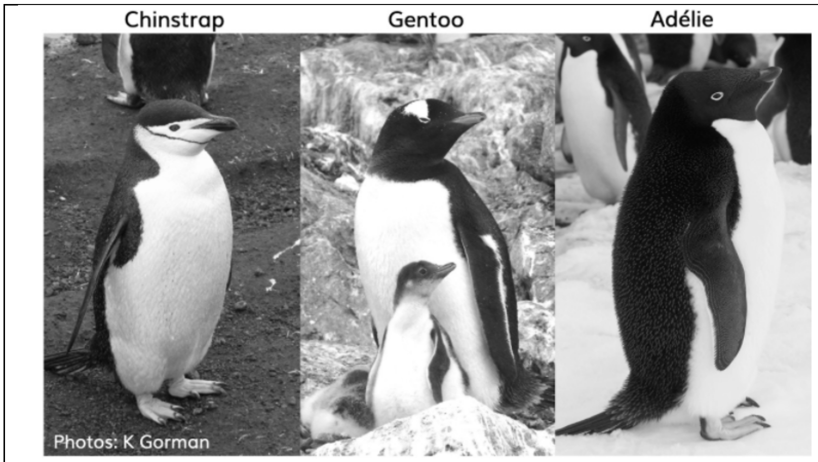
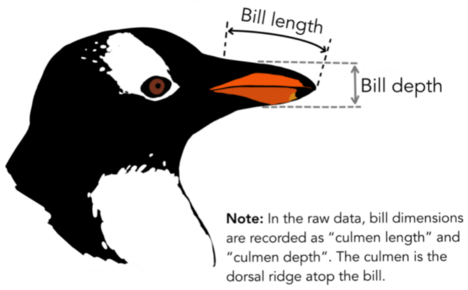


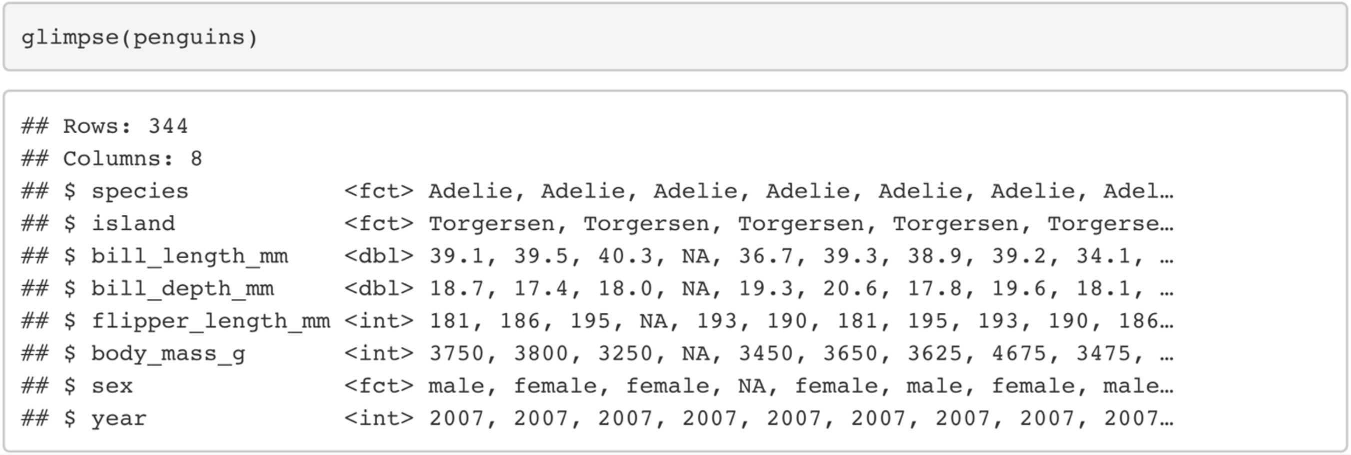
Table 3.3: Table of canonical data graphics and their corresponding **ggplot2** commands. Note that the mosaic plot function is not part of the **ggplot2** package.

response (y)	**explanatory (x)**	**plot type**	**geom_*()**
	numeric	histogram, density	`geom_histogram()` , `geom_density()`
	categorical	stacked bar	`geom_bar()`
numeric	numeric	scatter	`geom_point()`
numeric	categorical	box	`geom_boxplot()`
categorical	categorical	mosaic	`geom_mosaic()`

NOTE: We are going to make use the Palmer Penguins dataset in R. You can read more about the dataset at: https://education.rstudio.com/blog/2020/07/palmerpenguins-cran/. The dataset ‘penguins’ in the palmerpenguins package contains size measurements (such as bill length, bill depth, flipper length, and body mass), sex, and island for three penguin species observed on three islands in the Palmer Archipelago, Antarctica over a study period of three years.

 	 **Note:** In the raw data, bill dimensions are recorded as "culmen length" and "culmen depth". The culmen is the dorsal ridge atop the bill. Artwork by @allison_horst

EXAMPLE 3: Load the palmerpenguins package. (You may need to install this package first.) We can find some details about the dataset by using the glimpse() function.



```
glimpse(penguins)
```

```
## Rows: 344
## Columns: 8
## $ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adel…
## $ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgerse…
## $ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, …
## $ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, …
## $ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186…
## $ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475, …
## $ sex               <fct> male, female, female, NA, female, male, female, male…
## $ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007…
```

Layers of ggplot

NOTE: More advanced examples will follow, but ggplot function calls often start with the following foundation:

```
ggplot(data = [dataset],
          mapping = aes(x = [x-variable], y = [y-variable])) +
    geom_xxx() +
    other options
```

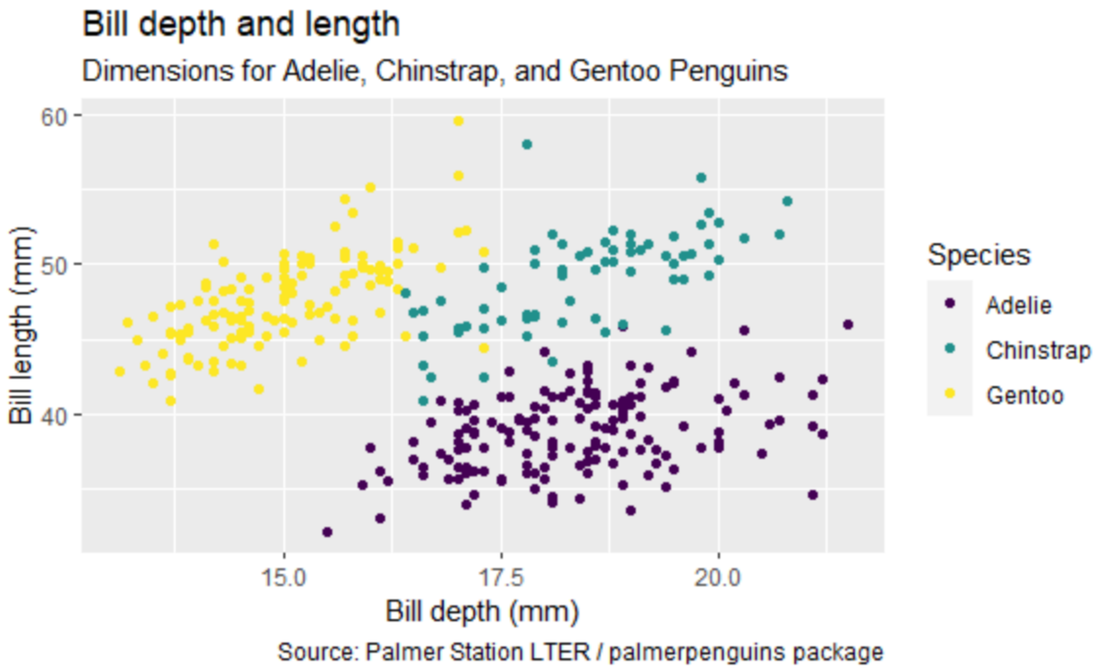
NOTE:

- ggplot() creates a coordinate system that you can add layers to.
- The first argument of ggplot() is the dataset to use in the graph. So ggplot(data = mtcars) creates an empty graph, but it's not very interesting
- The mapping argument defines how variables in your dataset are mapped to visual properties.
- The mapping argument is always paired with the aes() (short for aesthetic) function, and the x and y arguments of aes() specify which variables to map to the x and y axes.
- Aesthetics include details like the x/y axis variables and the size, shape, or color of your points.
- We complete the graph by adding one or more layers to ggplot(). For example, the geom_point() adds a layer of points to your plot, which creates a scatterplot.
- ggplot2 comes with many geom functions (such as geom_point, geom_boxplot, geom_histogram, etc.) that each add a different type of layer to a plot.
- A ggplot cheatsheet with many details, including various geom functions and their options, may be found at: https://ggplot2.tidyverse.org/

EXAMPLE 4: Create an appropriate plot for visualizing the bill length as a function of the bill depth. Add an aesthetic (here, use color) to incorporate the categorical variable species. Add appropriate titles.

The code shown below produces the plot. The goal is for you to understand and be able to write similar code:

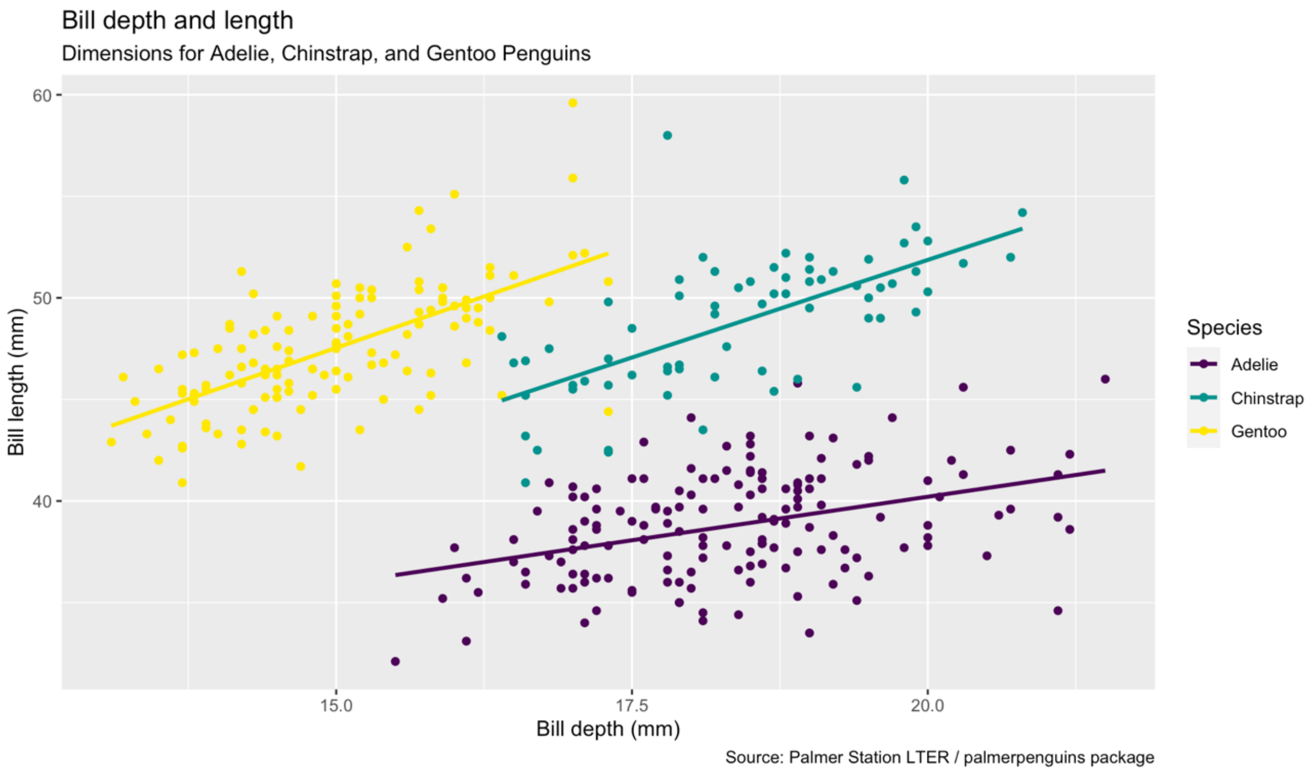
```
ggplot(data = penguins,
       mapping = aes(x = bill_depth_mm,
                     y = bill_length_mm,
                     color = species)) +
  geom_point() +
  labs(title = "Bill depth and length",
       subtitle = "Dimensions for Adelie, Chinstrap, and Gentoo Penguins",
       x = "Bill depth (mm)", y = "Bill length (mm)",
       color = "Species",
       caption = "Source: Palmer Station LTER / palmerpenguins package") +
  scale_colour_viridis_d()
```



NOTE: In order to fully understand how ggplot uses layers to build a plot, please see the slides found at: https://datasciencebox.org/course-materials/_slides/u2-d02-ggplot2/u2-d02-ggplot2.html#7. Stop when you reach the slide whose wepgage ends in ggplot2.html#19.

IDEA: The aesthetics are an important part of the ggplot functionality. Aside from mapping variables to the axes, aesthetics allow characteristics of glyphs (symbols on plot) to be mapped to a specific variable in the data. Examples aesthetics include: color (or colour), shape, size, alpha (alpha controls the transparency).

EXAMPLE 5: Let's add another layer to our plot. Write a command that will add a regression trend line for each species as shown below. What have you learned about the penguins based on this plot?

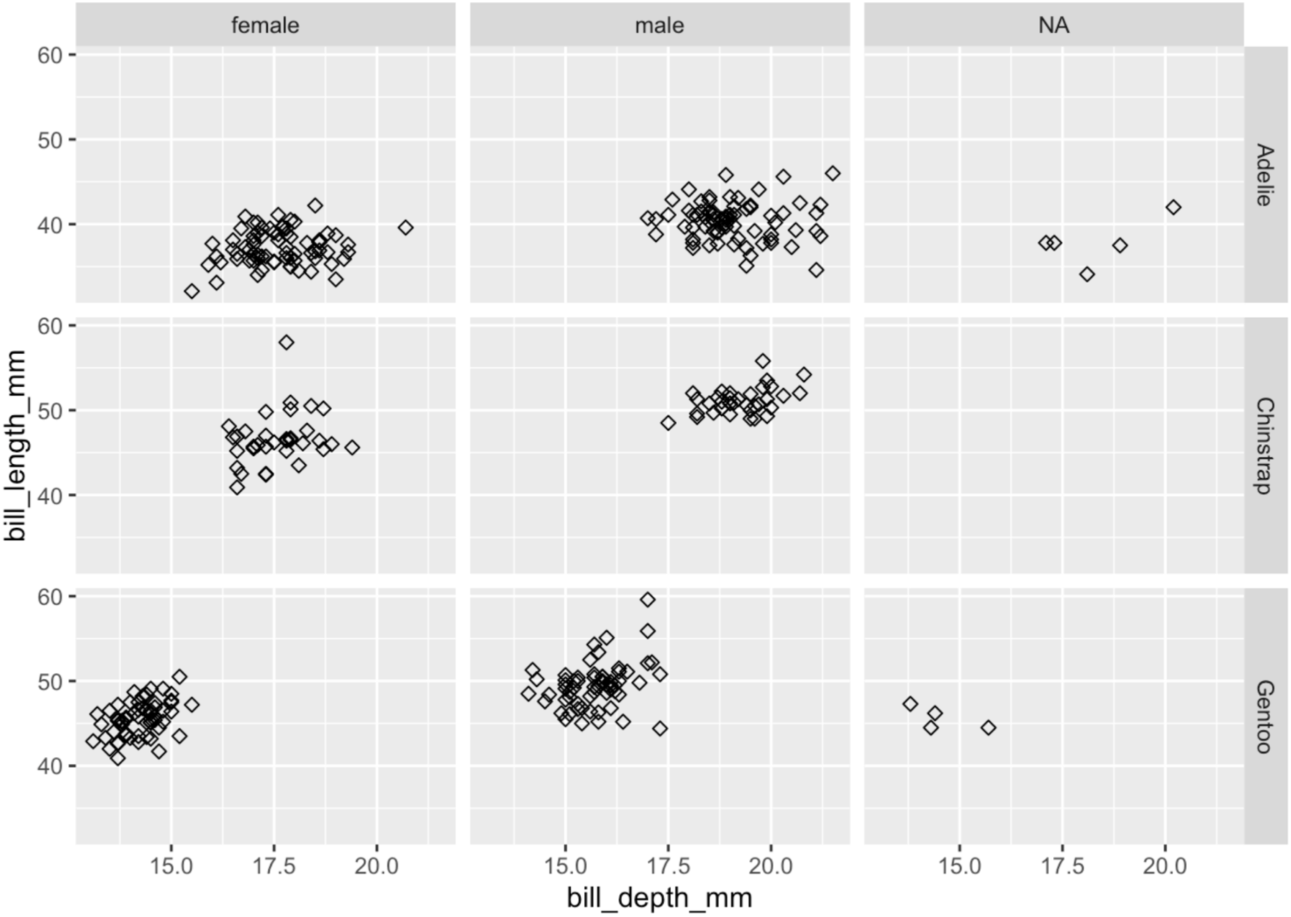


CAUTION: Using multiple aesthetics such as shape, color, and size to display multiple variables can produce a confusing, hard-to-read graph.

IDEA: Instead, facets—multiple side-by-side graphs used to display levels of a categorical variable—provide a simple and effective alternative.

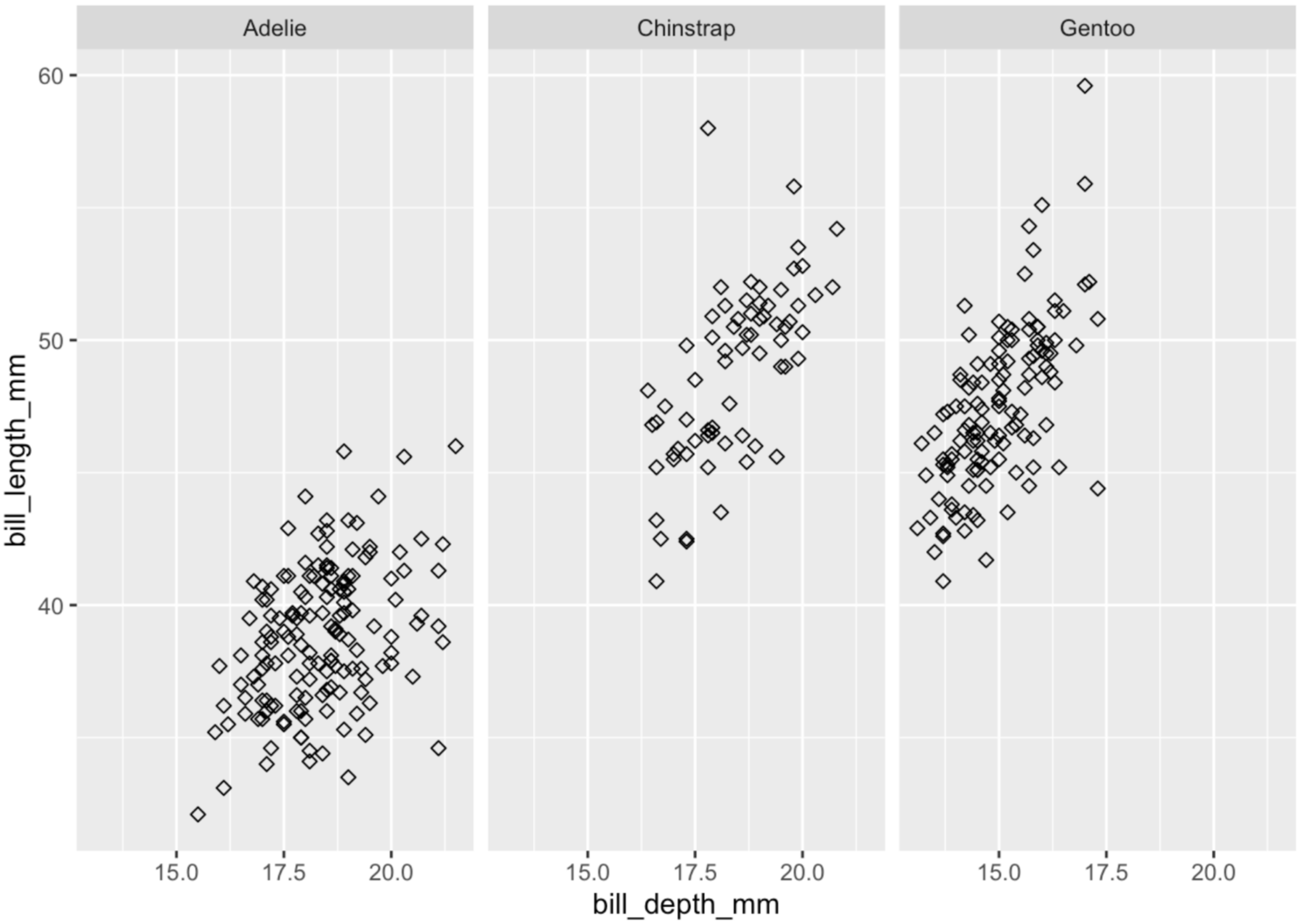
EXAMPLE 6: Run the following code to show the relationship between bill depth, bill length, sex, and species. To do this, we will facet on species and sex.

```
ggplot(penguins, aes(x = bill_depth_mm, y = bill_length_mm)) +
  geom_point(shape = 5) +
  facet_grid(species ~ sex)
```

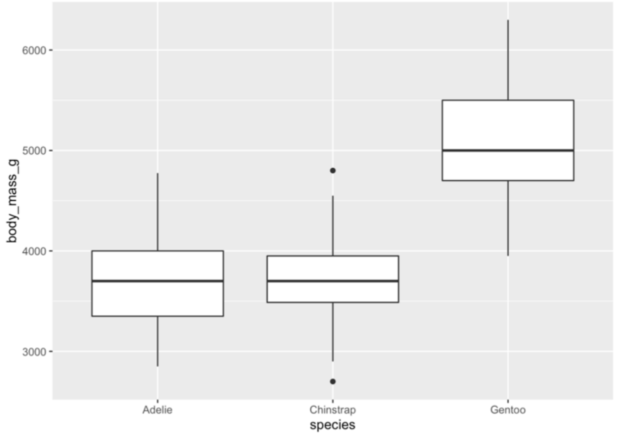
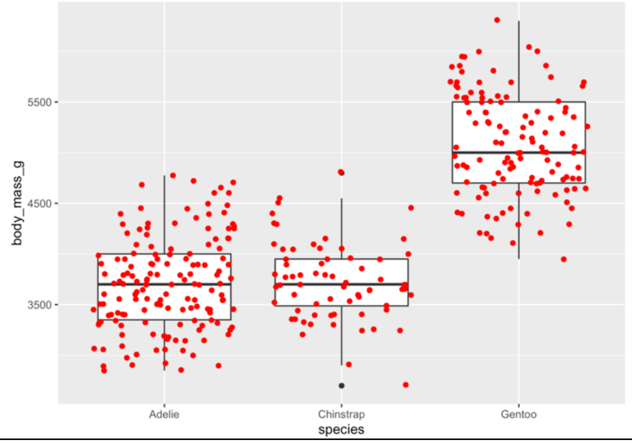


EXAMPLE 7: Repeat Example 6, but only facet on species to produce the plot shown below.

```
ggplot(penguins, aes(x = bill_depth_mm, y = bill_length_mm)) +
  geom_point(shape = 5) +
  facet_grid(species ~ .)
```



EXAMPLE 8: Not every plot is a scatterplot. Recreate the plots shown below. The second plot shows the boxplots, but also includes a depiction of the real data with "jitter".

 	
Plot A	Plot B