CS&SS Math Camp: Plotting in R

Emily A. Finchum-Mason

9/18/2020

Plotting in R

Data visualization is a powerful tool for understanding your data and teasing out relationships. This lab will illustrate how to visualize the distribution of variables as well as associations between variables using base `R` and the `tidyverse` plotting package.

For this lab, we will use two important packages: - `dplyr` to manage and manipulate our data - `ggplot2` to actually produce the visualizations

To see all of the cool data visualization that you can do in R, visit the R Graph Gallery: https://www.r-graph-gallery.com/

```
#install.packages("ggplot2")
#install.packages('palmerpenguins')
library(ggplot2)
library(dplyr)
library(palmerpenguins)
```

The data for this lab can be accessed through a package called `palmerpenguins`, which contains data on penguins' species, island, and body dimensions. This dataset will not actually appear in the global environment when you load the `palmerpenguins` library, but it is still accessible.

We will start by examining some descriptive statistics for each variable using the `summary()` command.

```
summary(penguins)
```

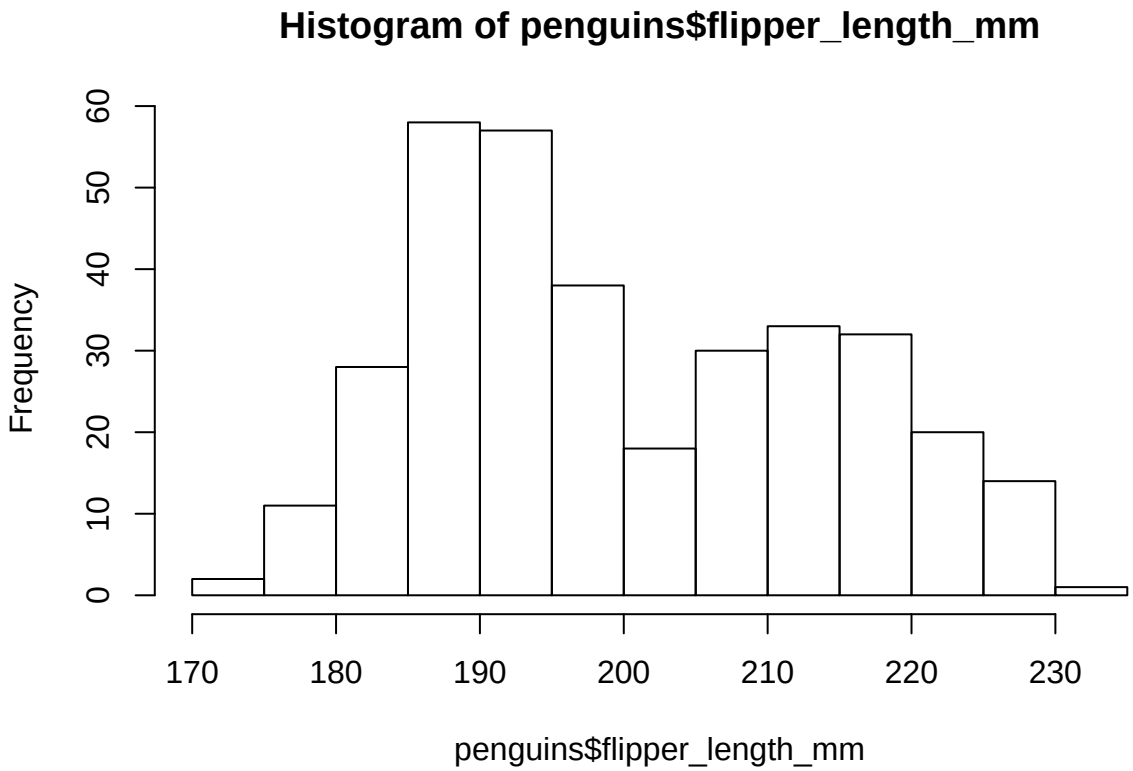
```
##       species          island    bill_length_mm  bill_depth_mm
##  Adelie   :152   Biscoe   :168   Min.   :32.10   Min.   :13.10
##  Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60
##  Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30
##                                  Mean   :43.92   Mean   :17.15
##                                  3rd Qu.:48.50   3rd Qu.:18.70
##                                  Max.   :59.60   Max.   :21.50
##                                  NA's   :2       NA's   :2
##  flipper_length_mm  body_mass_g       sex           year
##  Min.   :172.0     Min.   :2700   female:165   Min.   :2007
##  1st Qu.:190.0     1st Qu.:3550   male  :168   1st Qu.:2007
##  Median :197.0     Median :4050   NA's  : 11   Median :2008
##  Mean   :200.9     Mean   :4202                Mean   :2008
##  3rd Qu.:213.0     3rd Qu.:4750                3rd Qu.:2009
##  Max.   :231.0     Max.   :6300                Max.   :2009
##  NA's   :2         NA's   :2
```

Examining a single variable (univariate plots)

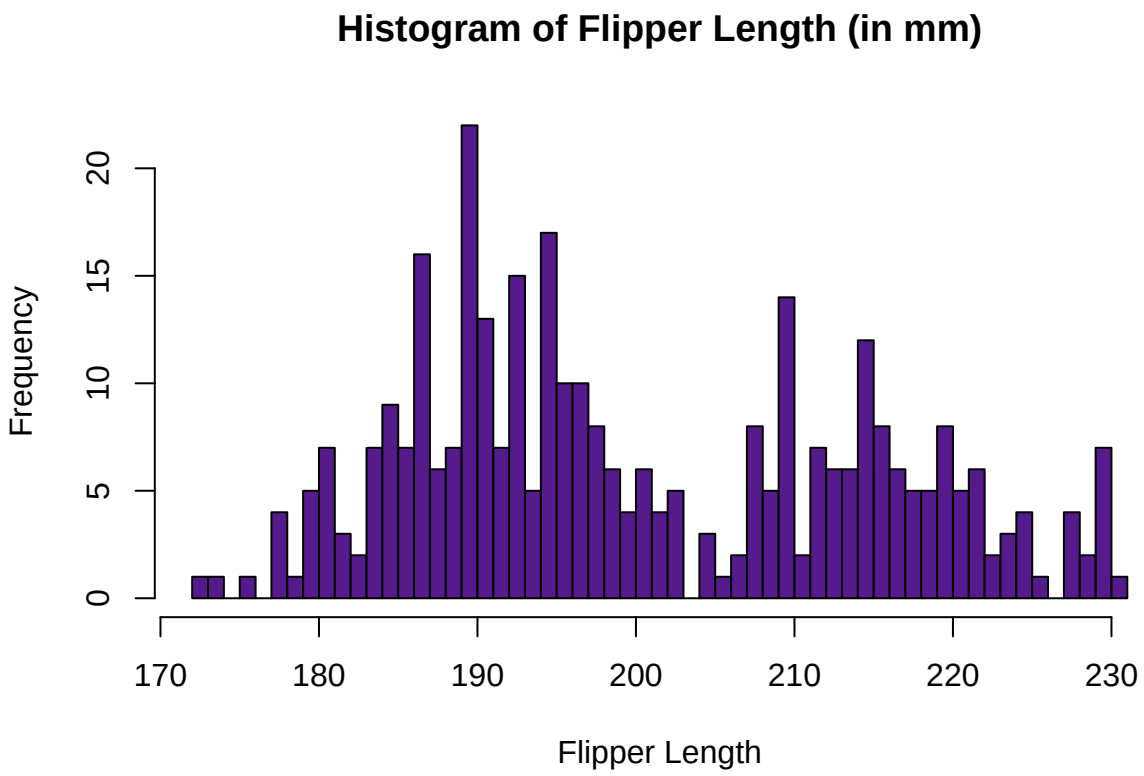
Oftentimes, we want to know how our quantitative variable is distributed in the data. Histograms provide handy way to do this. In base `R`, we simply use the `hist()` command. This is going to give us a very rough looking plot, but it's sufficient for exploratory data analysis (i.e. to get a sense of the typical case, whether and how the data is skewed).

Note that the `echo = FALSE` parameter in the code chunk to prevent printing of the R code that generated the plot.

Histograms using base R



If we were inclined to present this graph to an audience, we would want to clean it up, including changing titles, axis labels, the size of the "bins", and the color of the bars. For a more extensive list of the colors that you can choose, visit the following website: https://www.nceas.ucsb.edu/sites/default/files/2020-04/colorPaletteCheatsheet.pdf



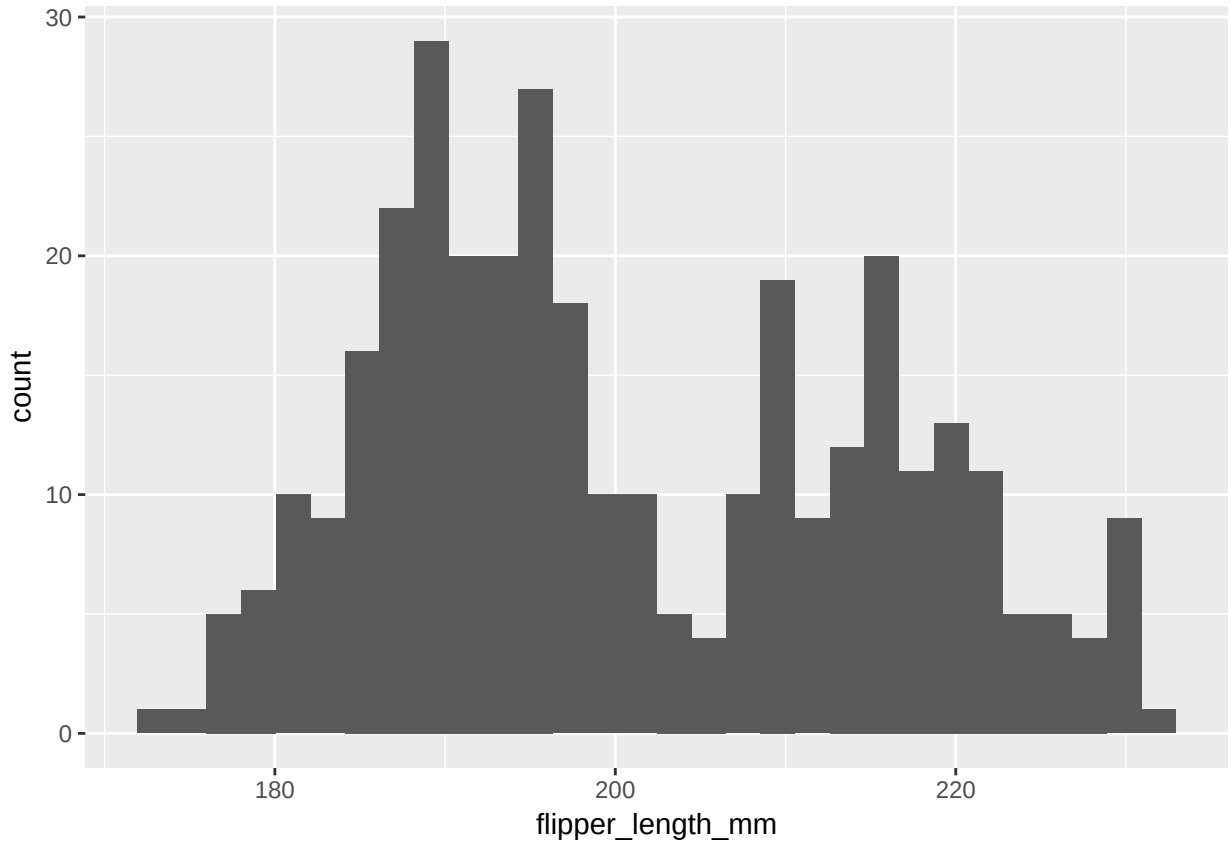
Question What do you notice happened when we changed the bin width in the second histogram? How might this be important to us as researchers?

Histograms in `ggplot2`

We can also use `ggplot2` to plot this variable. "gg" stands for "grammar of graphics"; the `ggplot2` package treats plots as an additive series of characteristics. The first line generally indicates (a) what dataset you are using and (b) what variable(s) will go into the aesthetics. Subsequent lines indicate the type of graph to be created (in this case, a histogram), different color attributes of the graph, axis titles and limits, et cetera.

The code in the first chunk is the simplest possible version of the histogram.

```
ggplot(penguins, aes(x = flipper_length_mm)) +
  geom_histogram()
```

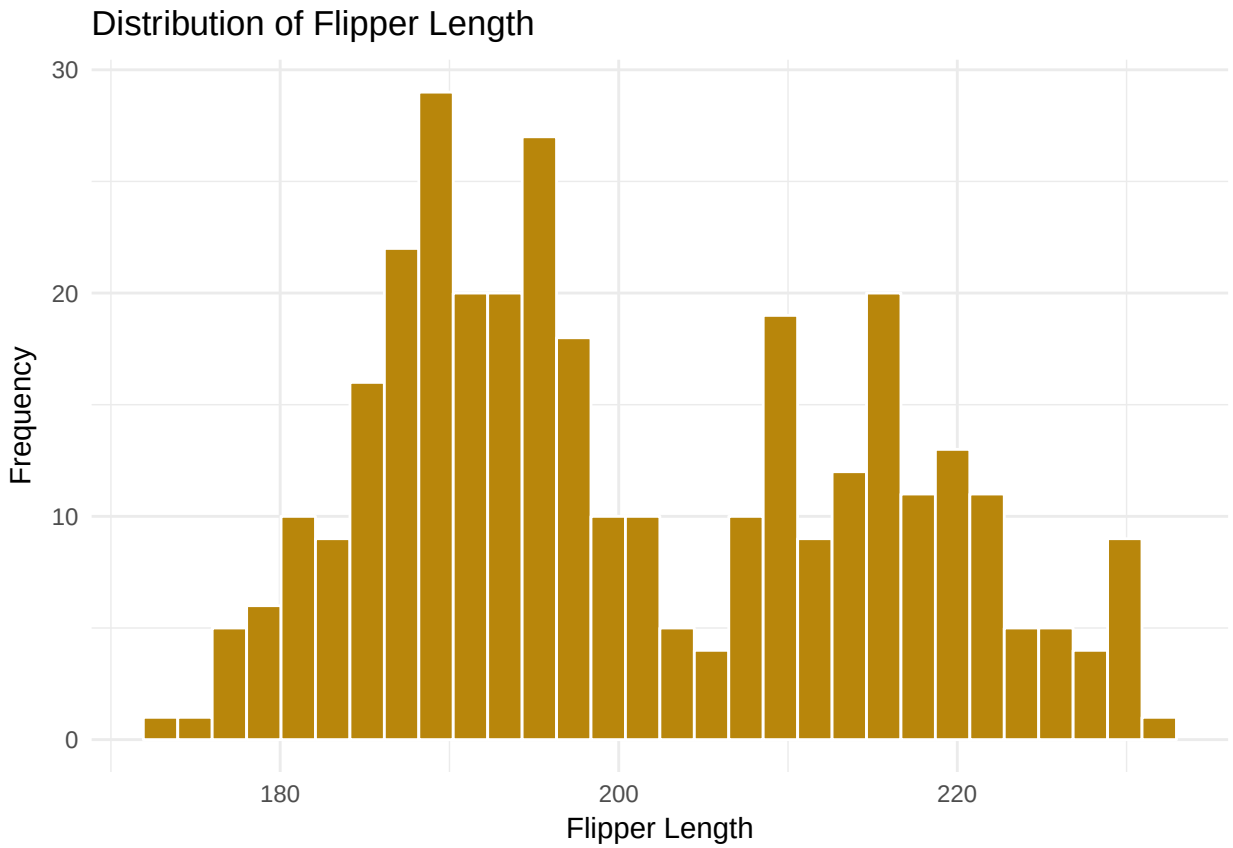


Again, it is not the most visually pleasing graph, but we can add features easily.

```
ggplot(penguins, aes(x = flipper_length_mm)) +
  geom_histogram(fill = "darkgoldenrod", #alter the fill color of the bars
                 color = "white") +      #alter the outline color of the bars
  ggtitle("Distribution of Flipper Length") +
  xlab("Flipper Length") +
  ylab("Frequency") +
  theme_minimal()     #changes the background theme.  The default is grey.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```

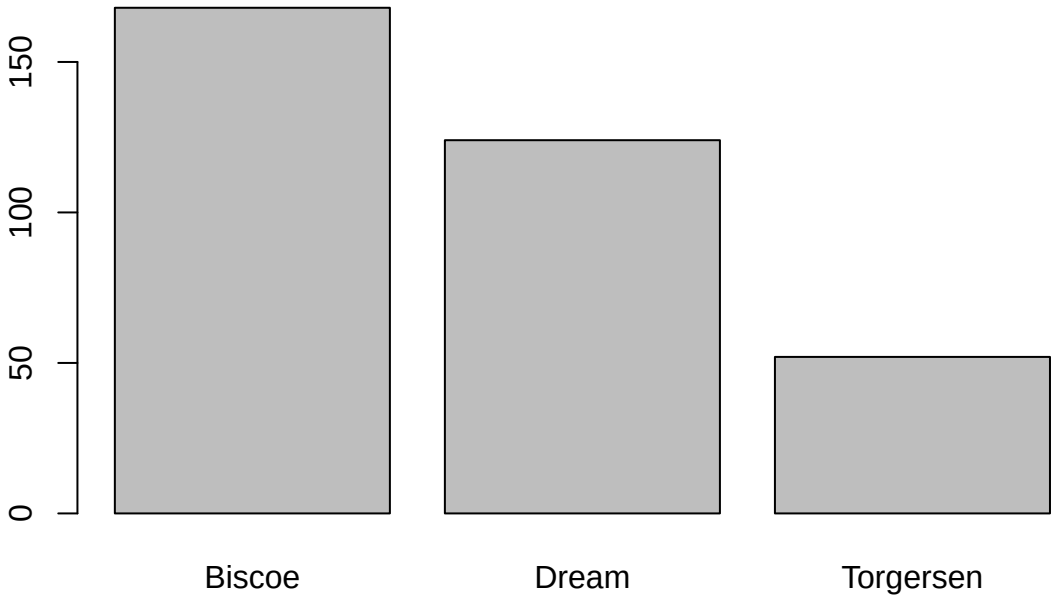


Bar graphs in R

Sometimes we're interested in visualizing a **categorical** variable In the case of the `penguins` dataset,

We'll first use base `R` and then `ggplot2` to create the bar plots. In base `R`, you must first convert your data into a table as we have done above.

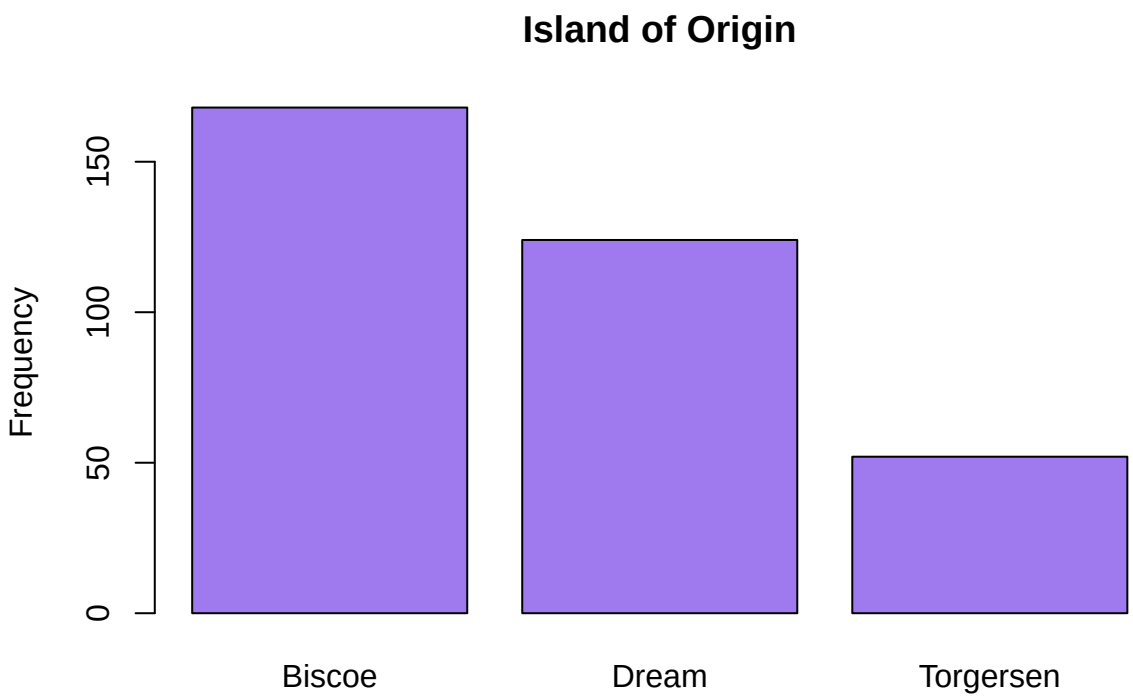
```
#Summarize # of penguins from each
#island in a table
counts <- table(penguins$island)

#Create the plot
barplot(counts)
```



This plot tells you how many observations (a.k.a. penguins) originated from each island.

We can add some bells and whistles to this plot in base `R` as well.

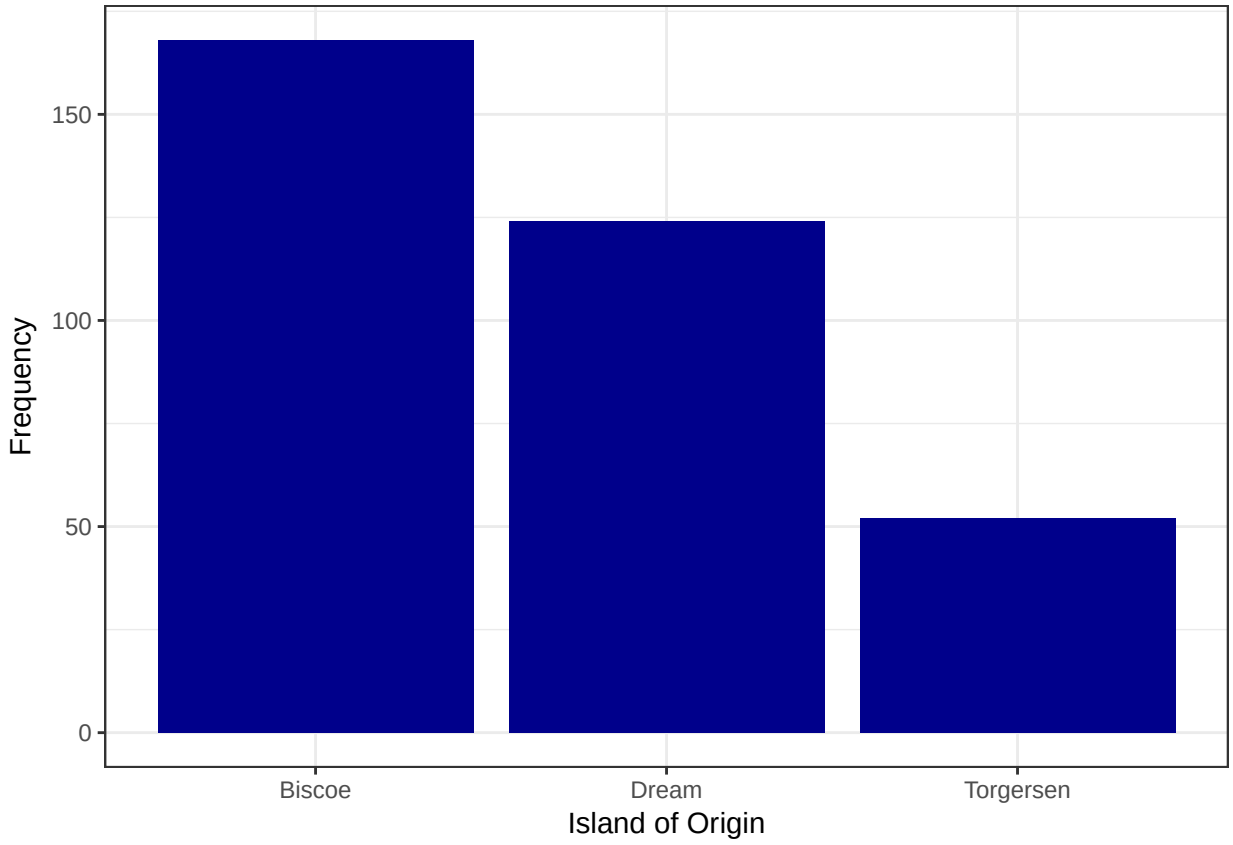
```
barplot(counts,
        main = "Island of Origin",
        col = "mediumpurple2",
        ylab = "Frequency")
```



In `ggplot2`, we will use the `geom_bar()` command. In `ggplot2`, we do not need to tabulate the data prior to plotting, which is handy.

```
ggplot(penguins, aes(x = island)) +
  geom_bar(fill = "darkblue") +
```

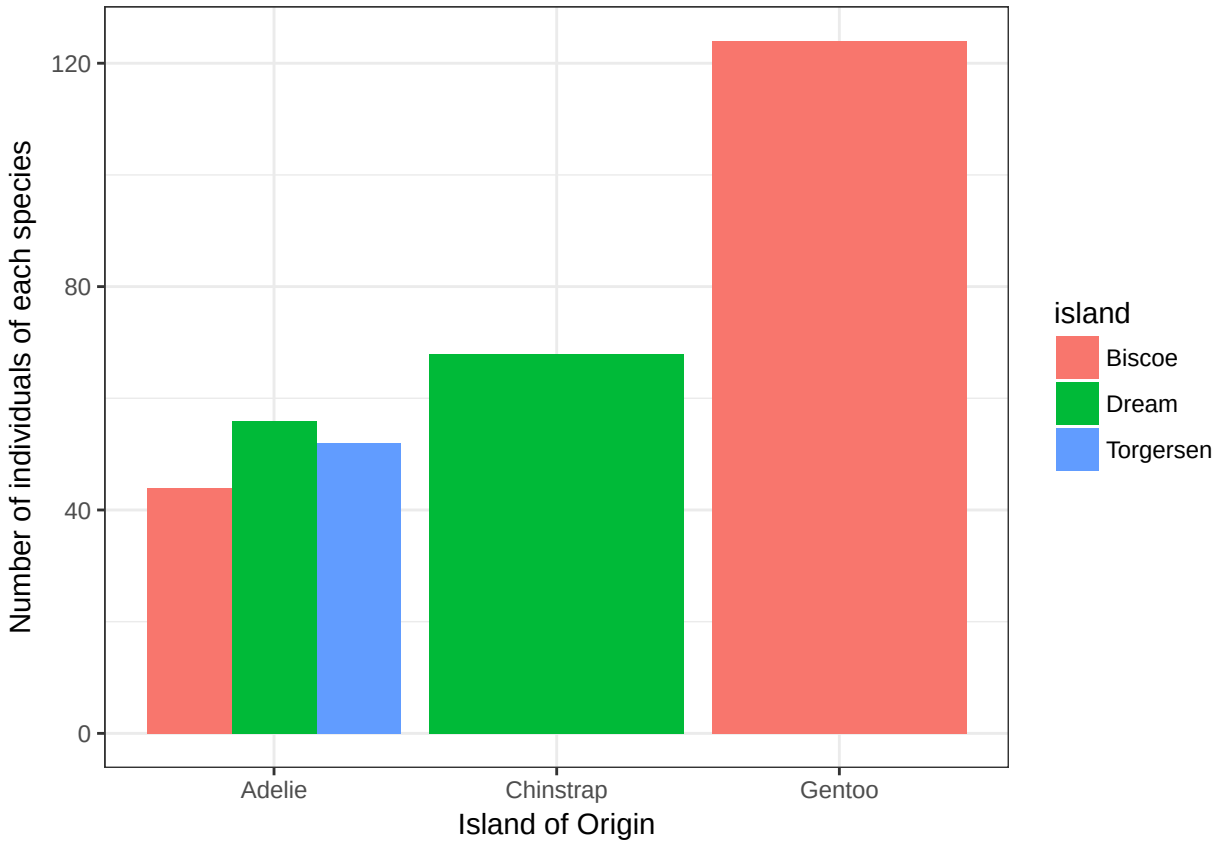
```
  xlab("Island of Origin") +
  ylab("Frequency") +
  theme_bw()    #note the difference in background with this theme
```



Grouped bar plots

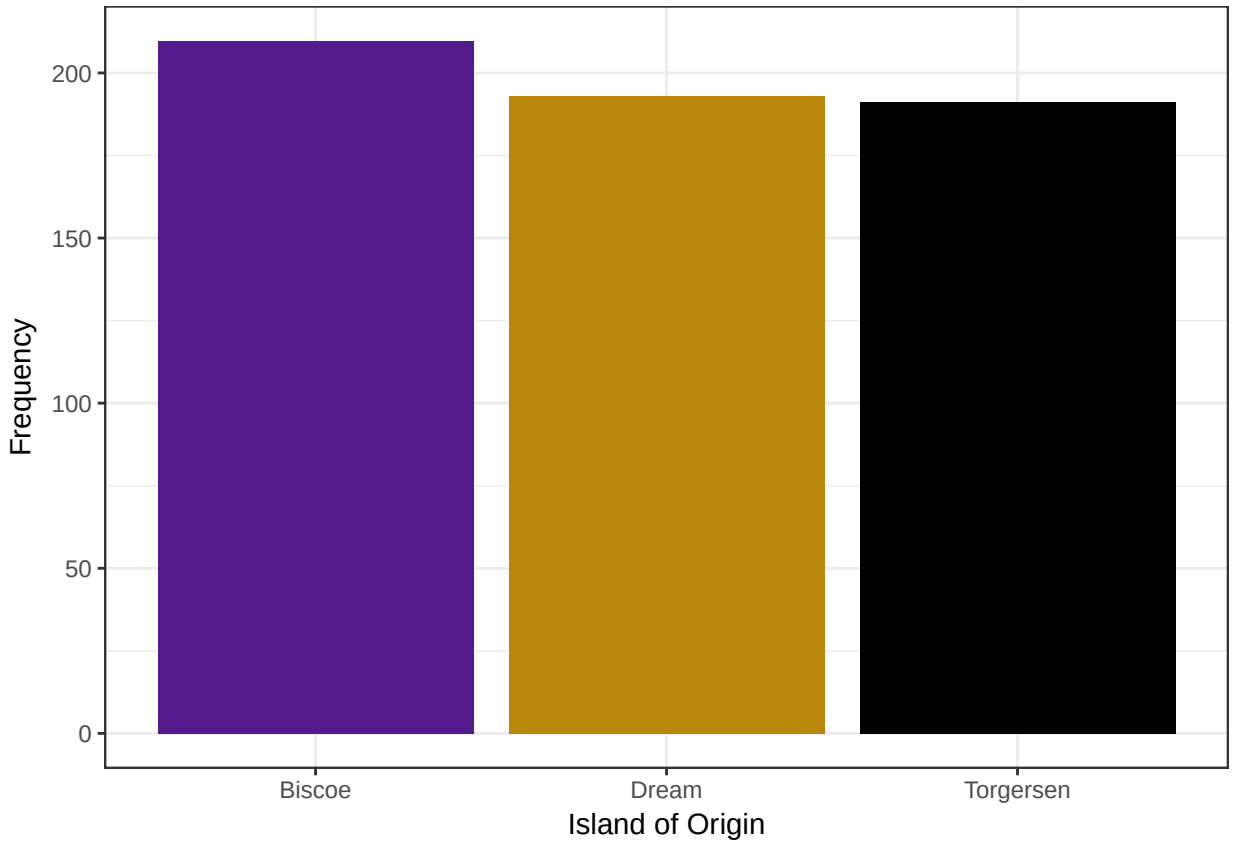
What if we want to observe the relationship between two categorical variables? In this case, we will examine the number of penguins of each species originating from each island. We can do so using a **grouped bar plot**. This is particularly straightforward to implement using `ggplot2`, but we will need to use `dplyr` to construct a summary tally of penguins by species and island and use that to make the graph.

```
#Make a quick summary table of penguins
#by island:
species_by_island <- penguins %>%
  group_by(species, island) %>%
  tally()

#Graph out the summary table using ggplot2:
ggplot(species_by_island, aes(x = species, y = n, fill = island)) +
  geom_bar(stat = "identity", position = "dodge") +
  xlab("Island of Origin") +
  ylab("Number of individuals of each species") +
  theme_bw()
```



We can also examine variable averages by group using a similar strategy as above. This time, we'll look at average flipper length by island. Again, we need to start by manipulating the data.

```
#summarize average flipper length by island
flippers_by_island <- penguins %>%
  group_by(island) %>%
  summarize(mean_flipper_length = mean(flipper_length_mm, na.rm = T))

#graph it out:
ggplot(flippers_by_island, aes(x = island, y = mean_flipper_length, fill = island)) +
  geom_bar(stat = "identity", position = "dodge") +
  xlab("Island of Origin") +
  ylab("Frequency") +
  theme_bw() +
  theme(legend.position = "none")  +  #use this last line to get rid of the legend
  scale_fill_manual(values = c("purple4", "darkgoldenrod", "black"))
```

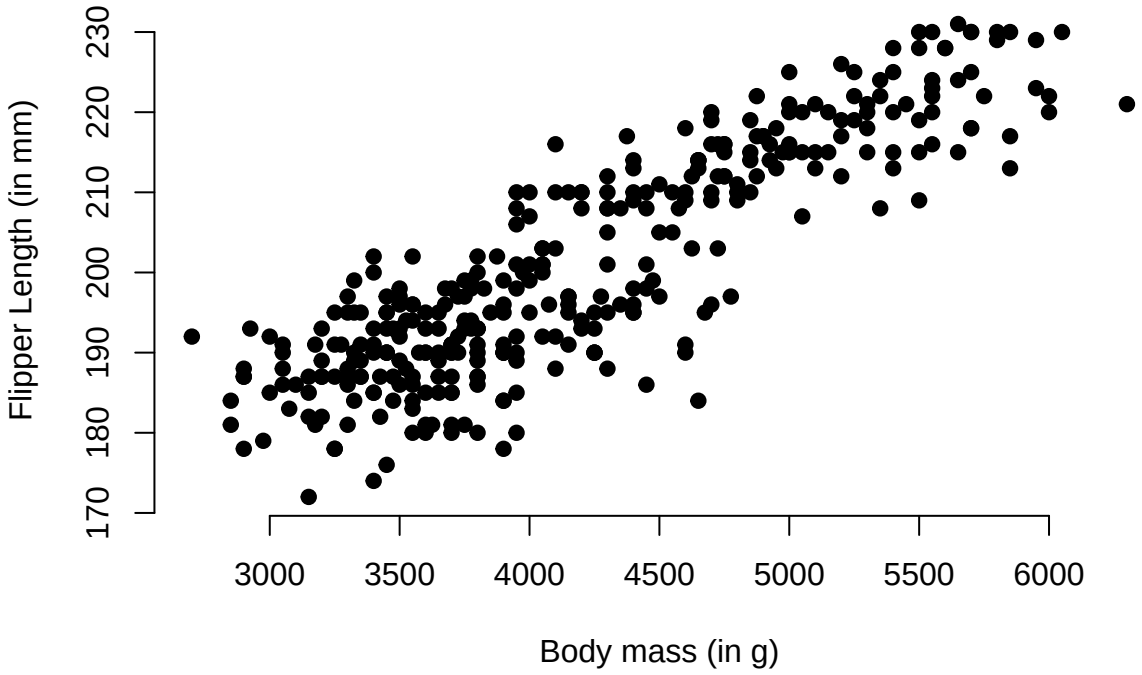


Scatterplots in R

To observe relationships between two numeric variables, we use scatterplots (also known as scattergrams). Say we want to observe the relationship between a penguin's body mass (`body_mass_g`) and the length of its flippers (`flipper_length_mm`).

Scatter plots in base R

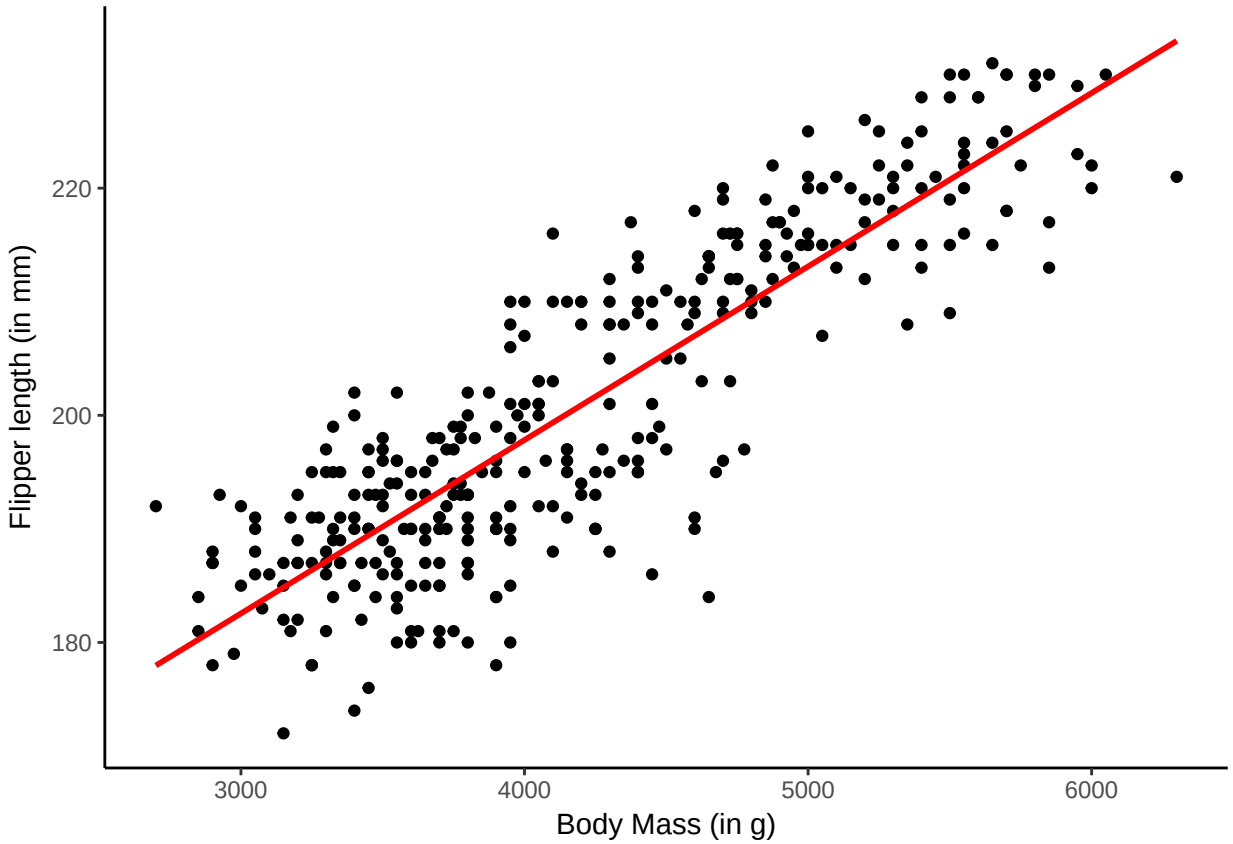
```
plot(penguins$body_mass_g, penguins$flipper_length_mm,
     xlab = "Body mass (in g)",
     ylab = "Flipper Length (in mm)",
     frame = FALSE, #removes the boundary that R automatically puts on a scatterplot
     pch = 19)  #specifies the type of dot
```



Scatterplots in ggplot2

We also might want to add a regression line to our scatterplot. While you can certainly do this in base `R`, the `ggplot2` method is given below.

```
ggplot(penguins, aes(x = body_mass_g, y = flipper_length_mm)) +
  geom_point() +
  xlab("Body Mass (in g)") +
  ylab("Flipper length (in mm)") +
  theme_classic() +
  geom_smooth(method = lm, color = "red", se=FALSE)
```

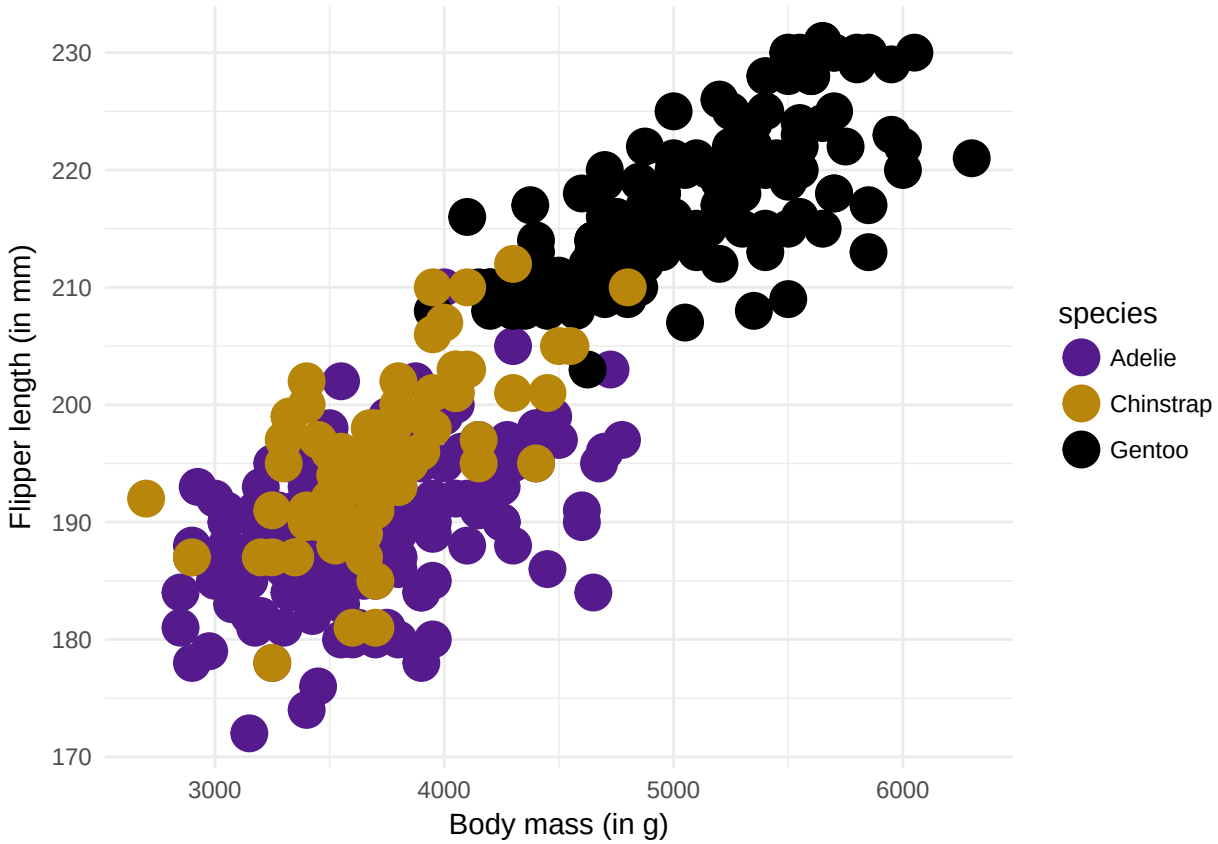



```
#lm = linear model
#se = whether or not to include standard error estimates
```

Grouped scatter plots

We can also observe how species might impact the relationship between body mass and flipper length using the grouping method. Here, we assign the dots of each species their own color. You can also distinguish categories using different levels of shading, different shapes, or even different sizes of points (https://www.r-graph-gallery.com/274-map-a-variable-to-ggplot2-scatterplot.html).

```
ggplot(penguins, aes(x=body_mass_g, y=flipper_length_mm, color=species)) +
    geom_point(size=6) +
    theme_minimal() +
  scale_color_manual(values = c("purple4", "darkgoldenrod", "black")) +
  xlab("Body mass (in g)") +
  ylab("Flipper length (in mm)")
```



This plot gives us a new perspective on the data: we see that the relationship between body mass and flipper length is quite different for the Gentoo species, as compared with the Adelie and Chinstrap species.

Plotting Exercises

Complete the following exercises for extra practice with plotting in R. Feel free to use either base `R` or `ggplot2` (or experiment with both).

Construct your graphs and answer the associated questions in a fresh `R Markdown` file.

These exercises will require the `mpg` dataset. To get the cars dataset to load, simply enter `data("mpg")` into a new chunk. You can use the `summary(mpg)` command or the `View(mpg)` command to see what is in the dataset.

1. Using the `mpg` dataset, create a histogram of the highway mileage (`hwy`) variable. Describe what you observe about the distribution of this variable.
2. Create a bar plot for the class of vehicles (`class`). What is the most common class of car observed in this data? What is the least common class of car observed in this data?
3. How does average city gas mileage (`cty`) vary by transmission (`trans`) type? Create a graph that offers some insight to this question and describe what you observe.
4. What is the relationship between engine displacement (`displ`) and highway gas mileage (`hwy`)? Create a graph that offers some insight to this question and describe what you observe.
5. How does the relationship between engine displacement and highway gas mileage vary by transmission type? Create a graph that offers some insight to this question and describe what you observe.