Exploratory Data Analysis

STA 610 - Applied Statistics for Health Professions

Andrew DiLernia

## Learning Objectives

* Create and interpret univariate statistics & visualizations
* Create and interpret bivariate statistics & visualizations

Let’s begin by loading some R packages for this activity using the code below. Note: if it is the first time you are using an R package, you may need to install it first using the install.packages() function.

*# Load necessary packages*  
library(tidyverse)  
library(ggthemes)  
library(flextable)  
library(corrr)

Next, we can set default theme settings for plots, and load some functions to simplify table customization and creation using the code below.

*# Set ggplot theme for visualizations*  
theme\_set(ggthemes::theme\_few())  
  
*# Set options for flextables*  
set\_flextable\_defaults(na\_str = "NA")  
  
*# Load function for printing tables nicely*  
source("https://raw.githubusercontent.com/dilernia/STA323/main/Functions/make\_flex.R")

In this activity we will analyze the Palmer Penguins data set: a data set consisting of measurements collected on 344 penguins from 3 islands in [Palmer Archipelago, Antarctica](https://www.google.com/maps/place/Palmer+Archipelago/@-64.1463814,-62.1800963,7z/data=!3m1!4b1!4m5!3m4!1s0xbc78dd6dc38c572b:0xe609367aeed33087!8m2!3d-64.1957848!4d-62.0153384). Specifically, data were collected and made available by [Dr. Kristen Gorman](https://www.uaf.edu/cfos/people/faculty/detail/kristen-gorman.php) and the [Palmer Station, Antarctica Long Term Ecological Research Network](https://pallter.marine.rutgers.edu/).

|  |  |
| --- | --- |

*Artwork by @allison\_horst*

Let’s import data from the website GitHub to use for this activity.

*# Load Palmer penguins data*  
penguins <- readr::read\_csv("https://raw.githubusercontent.com/dilernia/STA323/main/Data/penguins.csv")

*Data dictionary for Palmer Penguins data set.*

| Variable | Description |
| --- | --- |
| species | Species of the penguin |
| island | Island the penguin was found on |
| bill\_length\_mm | Bill length (mm) |
| bill\_depth\_mm | Bill depth (mm) |
| flipper\_length\_mm | Flipper length (mm) |
| body\_mass\_g | Body mass (g) |
| sex | Sex of the penguin |
| year | Year data was collected |

## Exploratory Data Analysis (EDA)

**Exploratory data analysis (EDA)** is the process of analyzing data to explore key characteristics, commonly using data visualizations and descriptive statistics.

Conducting EDA is contingent upon one’s data already being cleaned, organized in a tidy format (i.e., variables are columns and observations are rows), and imported into R or another software tool so that it is ready for analysis. We will not be covering the process of data cleaning or importing data into R, but the [R for Data Science (2e)](https://r4ds.hadley.nz/) text is a great resource for learning more about these steps.

EDA is an essential skill for being a statistician or data scientist, with its main purposes being:

* Identification of outliers
* Exploring patterns of missing data
* Formulating hypotheses for studies based on new data
* To generally improve our understanding of the data

There are two main components of EDA: descriptive statistics and data visualizations.

## Descriptive statistics

**Descriptive statistics** are numerical measures describing the *sample*, a collection of individuals or entities we have measurements for. Descriptive statistics are different from inferential statistics which involve the *population* of interest, the full set of individuals or entities one desires to study or conduct inference about.

## Data visualization

**Data visualization** is the process of graphically representing our data. The main goals of data visualization are to:

* Efficiently present data in a compact space
* Make data sets coherent / understandable to our target audience
* Serve a clear purpose
* Not distort what data has to say

For a more detailed set of guidelines regarding best practices of data visualization, see this [guide](https://drive.google.com/file/d/1BPZcdrFgoj_TjzLVWUImZ92gXCKuaGy6/view?usp=share_link) created by people working at the pharmaceutical company Novartis.

## Univariate analyses

### Univariate descriptive statistics

For a single quantitative variable, some commonly used univariate statistics are listed below.

#### Measures of location

* **Mean**: average value,
* **Minimum**: smallest value,
* **First quartile**: 25th percentile, lower-quartile, Q1
* **Median**: 50th percentile, Q2
* **Third quartile**: 75th percentile, upper-quartile, Q3
* **Maximum**: largest value,

For a sample with observations, the sample mean is

The mean gives the average value of a variable, while the median is the number such that of values are greater than or equal to this value, and are less than or equal to this value. The mean is sensitive to extreme or outlier observations, while the median is robust (less sensitive) to extreme values / outliers, so the median is a more appropriate measure of center when outliers are present.

#### Measures of spread

* **Range**:
* **Interquartile Range (IQR)**:
* **Standard deviation**:
* **Variance**:

Measures of variability or dispersion describe how spread out values are for a particular variable.

For a sample with observations, the formula for the sample variance is

### Five-Number Summary

The five-number summary is a commonly used set of statistics, consisting of the , Q1, the median, Q3, and .

Let’s look at some statistics for the flipper lengths in mm of the Palmer penguins.

*# Calculating descriptive statistics*  
quant1Stats <- penguins %>%   
 dplyr::summarize(  
 Minimum = min(flipper\_length\_mm, na.rm = TRUE),  
 Q1 = quantile(flipper\_length\_mm, na.rm = TRUE, probs = 0.25),  
 M = median(flipper\_length\_mm, na.rm = TRUE),  
 Q3 = quantile(flipper\_length\_mm, na.rm = TRUE, probs = 0.75),  
 Maximum = max(flipper\_length\_mm, na.rm = TRUE),  
 Mean = mean(flipper\_length\_mm, na.rm = TRUE),  
 R = Maximum - Minimum,  
 s = sd(flipper\_length\_mm, na.rm = TRUE)  
)  
  
*# Printing table of statistics*  
quant1Stats %>%   
 make\_flex(caption = "Quantitative summary statistics for penguin flipper lengths (mm).")

*Table 1: Quantitative summary statistics for penguin flipper lengths (mm).*

| Minimum | Q1 | M | Q3 | Maximum | Mean | R | s |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 172.00 | 190.00 | 197.00 | 213.00 | 231.00 | 200.92 | 59.00 | 14.06 |

Note that this data set contains missing values, denoted as NA in R, but these were excluded when calculating the sample statistics above. This is how we will handle missing data moving forward.

Missing data are important to be aware of as their source should be investigated or questioned. Although we won’t be covering methods for handling missing data, information regarding this can be found in Chapter 17 of [Regression and Other Stories](https://users.aalto.fi/~ave/ROS.pdf) by Gelman, Hill, & Vehtari (2020).

➡️ What are the largest and smallest flippers lengths for penguins in this data set?

➡️ Provide and interpret the value of the sample median flipper length for the penguins.

➡️ Provide the value of the sample variance of the flipper length for the penguins.

➡️ Which statistic is more sensitive to outliers: the range or the interquartile-range (IQR)?

➡️ **Try on your own**: Consider the set of values given by . Confirm the values of the following sample statistics.

*Quantitative summary statistics for example data set.*

| Minimum | Q1 | M | Q3 | Maximum | Var | SD | R | IQR | n | Mean |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1.00 | 3.00 | 4.00 | 9.00 | 9.00 | 13.20 | 3.63 | 8.00 | 6.00 | 5 | 5.20 |

Frequencies also known as counts are useful descriptive statistics for categorical variables. The categorical variables in this data set we can calculate frequency tables for are: species, island, and sex.

*# Printing frequency tables*  
penguins %>%   
 dplyr::count(species) %>%   
 make\_flex(caption = "Number of penguins by species.")

*Table 2: Number of penguins by species.*

| species | n |
| --- | --- |
| Adelie | 152 |
| Chinstrap | 68 |
| Gentoo | 124 |

➡️ You try: Recreate the frequency table for island shown below.

*Table 3: Number of penguins by island.*

| island | n |
| --- | --- |
| Biscoe | 168 |
| Dream | 124 |
| Torgersen | 52 |

➡️ Which penguin species had the most observations in this data set?

➡️ Which island had the least penguins measured on it?

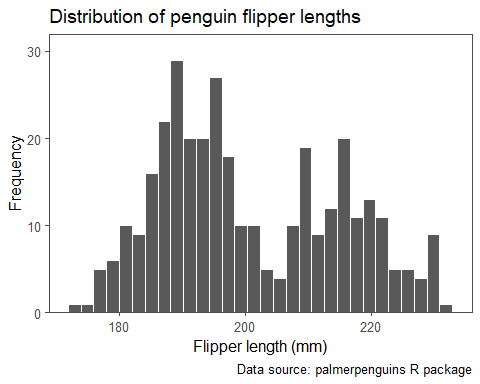
### Univariate visualizations

Data visualization is a key aspect of EDA. Univariate data visualizations are useful for understanding characteristics of a single variable in an efficient manner and also to identify potential outliers or unusual observations.

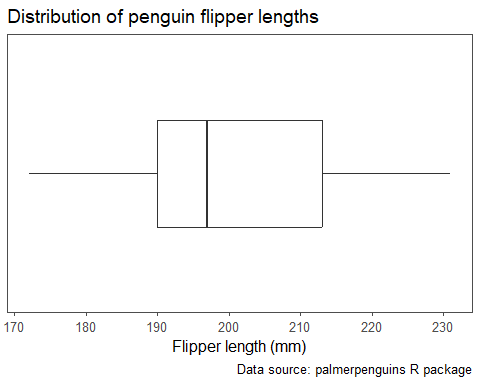
For a single quantitative variable, two commonly used data visualizations are the box plot and histogram.

➡️ Recreate the histogram and a box plot for the penguin flipper lengths using the code below.

*# Creating a histogram*  
penguins %>%   
 ggplot(aes(x = flipper\_length\_mm)) +   
 geom\_histogram(color = "white") +  
 scale\_y\_continuous(expand = expansion(mult = c(0, 0.10))) +  
 labs(title = "Distribution of penguin flipper lengths",  
 x = "Flipper length (mm)",  
 y = "Frequency",  
 caption = "Data source: palmerpenguins R package")



*# Creating a box plot*  
penguins %>%   
 ggplot(aes(x = flipper\_length\_mm)) +   
 geom\_boxplot() +  
 scale\_y\_discrete(breaks = NULL) +  
 labs(title = "Distribution of penguin flipper lengths",  
 x = "Flipper length (mm)",  
 caption = "Data source: palmerpenguins R package")



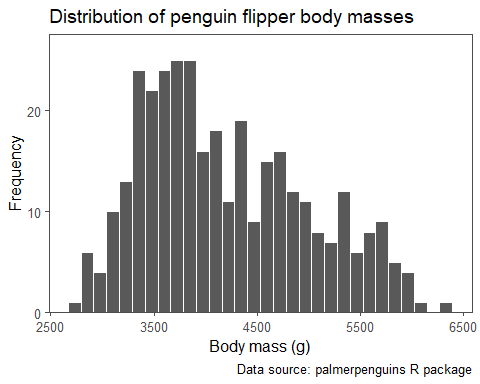
➡️ What can we say about the penguin flipper lengths based on the histogram?

The histogram shows that the distribution of flipper lengths is bimodal and fairly symmetric.

➡️ What can we say about the penguin flipper lengths based on the box plot? Are there any outliers present?

Box plot cannot say about the modal type.

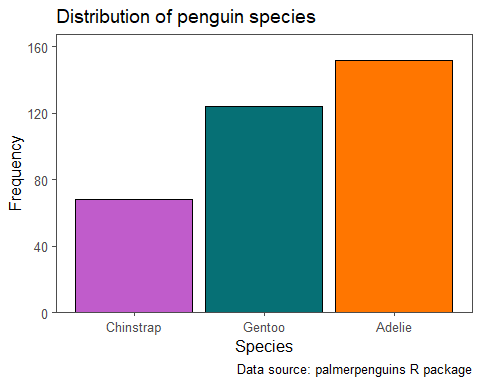
➡️ Modifying the code for the previous histogram, recreate the histogram below for the body mass variable.



For a single categorical variable, a commonly used data visualization is the bar chart.

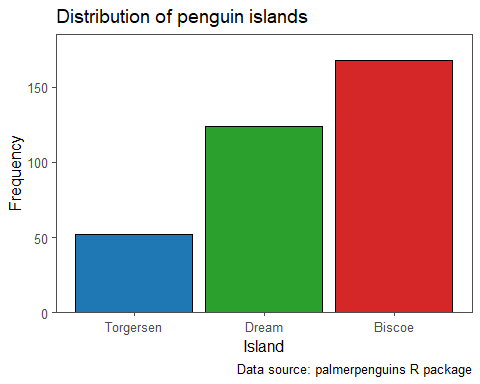
➡️ Recreate the bar chart displaying the number of penguins of each species below.

*# Creating a bar chart*  
penguins %>% dplyr::count(species, .drop = FALSE) %>%   
 mutate(species = fct\_reorder(species, n)) %>%   
 ggplot(aes(x = species, y = n,  
 fill = species)) +   
 geom\_col(color = "black") +  
 scale\_fill\_manual(values = c("#c05ccb", "#067075", "#ff7600")) +  
 scale\_y\_continuous(expand = expansion(mult = c(0, 0.10))) +  
 labs(title = "Distribution of penguin species",  
 x = "Species",  
 y = "Frequency",  
 caption = "Data source: palmerpenguins R package") +  
 theme(legend.position = "none")



➡️ What can we say about the penguin species based on the bar chart?

➡️ Modifying the code for the previous bar chart, recreate the bar chart below for the island variable. *Hint*: the hex codes for the color-blind friendly colors in the bar chart are c("#1F77B4", "#2CA02C", "#D62728").



## Bivariate analyses

### Bivariate descriptive statistics

For two quantitative variables, the sample correlation coefficient, , is a commonly used statistic which describes both the strength and direction of a linear relationship between the two variables. Correlations range from -1 to 1, with values closer to -1 indicating a stronger, negative linear relationship (as one variable increases, the other decreases), values closer to 0 indicating a weaker linear relationship, and values closer to 1 indicating a stronger, positive and linear relationship (as one variable increases, the other increases).

Correlations are a useful starting point, but should be used with caution as they have a few limitations:

1. They only measure linear relationships, but many relationships are non-linear. ⎰
2. A strong correlation does not imply causation, e.g., monthly ice cream sales may be strongly correlated with the number of monthly shark attacks, but that does not necessarily mean one causes the other. 🍦🦈
3. Confounding variables can distort the relationship between variables of interest.
4. The correlation coefficient is sensitive to outliers.

Let’s look at the correlations for the penguins data:

*# Calculating correlations*  
corTable <- penguins %>%   
 corrr::correlate(diagonal = 1)  
  
*# Printing table of correlations*  
corTable %>%  
 make\_flex(caption = "Table of pairwise correlations.")

*Table 4: Table of pairwise correlations.*

| term | bill\_length\_mm | bill\_depth\_mm | flipper\_length\_mm | body\_mass\_g | year |
| --- | --- | --- | --- | --- | --- |
| bill\_length\_mm | 1.00 | -0.24 | 0.66 | 0.60 | 0.05 |
| bill\_depth\_mm | -0.24 | 1.00 | -0.58 | -0.47 | -0.06 |
| flipper\_length\_mm | 0.66 | -0.58 | 1.00 | 0.87 | 0.17 |
| body\_mass\_g | 0.60 | -0.47 | 0.87 | 1.00 | 0.04 |
| year | 0.05 | -0.06 | 0.17 | 0.04 | 1.00 |

➡️ Which variables have the strongest correlation?

➡️ Which variables have the weakest correlation?

For one quantitative and one categorical variable, we can look at the same univariate statistics we have previously, but stratified by the groups of the categorical variable.

➡️ Recreate the table of statistics for the flipper lengths of the penguins stratified by species below.

*# Calculating descriptive statistics*  
quant2Stats <- penguins %>%   
 group\_by(species) %>%   
 summarize(  
 Minimum = min(flipper\_length\_mm, na.rm = TRUE),  
 Q1 = quantile(flipper\_length\_mm, na.rm = TRUE, probs = 0.25),  
 M = median(flipper\_length\_mm, na.rm = TRUE),  
 Q3 = quantile(flipper\_length\_mm, na.rm = TRUE, probs = 0.75),  
 Maximum = max(flipper\_length\_mm, na.rm = TRUE),  
 Mean = mean(flipper\_length\_mm, na.rm = TRUE),  
 R = Maximum - Minimum,  
 s = sd(flipper\_length\_mm, na.rm = TRUE),  
 n = n()  
)  
  
*# Printing table of statistics*  
quant2Stats %>%   
 make\_flex(caption = "Summary statistics for penguin flipper lengths by species.")

*Table 5: Summary statistics for penguin flipper lengths by species.*

| species | Minimum | Q1 | M | Q3 | Maximum | Mean | R | s | n |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Adelie | 172.00 | 186.00 | 190.00 | 195.00 | 210.00 | 189.95 | 38.00 | 6.54 | 152 |
| Chinstrap | 178.00 | 191.00 | 196.00 | 201.00 | 212.00 | 195.82 | 34.00 | 7.13 | 68 |
| Gentoo | 203.00 | 212.00 | 216.00 | 221.00 | 231.00 | 217.19 | 28.00 | 6.48 | 124 |

➡️ Which penguin species typically had the largest flipper lengths?

➡️ Which penguin species had the most variability in their flipper lengths?

➡️ Modifying the previously provided code, recreate the table of statistics for the body masses of the penguins stratified by sex below.

*Table 6: Summary statistics for penguin body masses by sex.*

| sex | Minimum | Q1 | M | Q3 | Maximum | Mean | R | s | n |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| female | 2,700.00 | 3,350.00 | 3,650.00 | 4,550.00 | 5,200.00 | 3,862.27 | 2,500.00 | 666.17 | 165 |
| male | 3,250.00 | 3,900.00 | 4,300.00 | 5,312.50 | 6,300.00 | 4,545.68 | 3,050.00 | 787.63 | 168 |
| NA | 2,975.00 | 3,475.00 | 4,100.00 | 4,650.00 | 4,875.00 | 4,005.56 | 1,900.00 | 679.36 | 11 |

For two categorical variables, a table of counts is commonly calculated as below.

*# Creating frequency table*  
speciesIslandCounts <- penguins %>%   
 dplyr::count(species, island)  
  
*# Printing frequency table*  
speciesIslandCounts %>%   
 make\_flex(caption = "Number of penguins by island and species.")

*Table 7: Number of penguins by island and species.*

| species | island | n |
| --- | --- | --- |
| Adelie | Biscoe | 44 |
| Adelie | Dream | 56 |
| Adelie | Torgersen | 52 |
| Chinstrap | Dream | 68 |
| Gentoo | Biscoe | 124 |

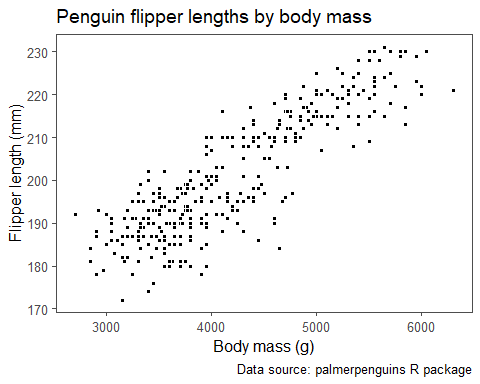
➡️ Was any penguin species found on more than one island?

➡️ How many penguins were found on Dream island?

### Bivariate visualizations

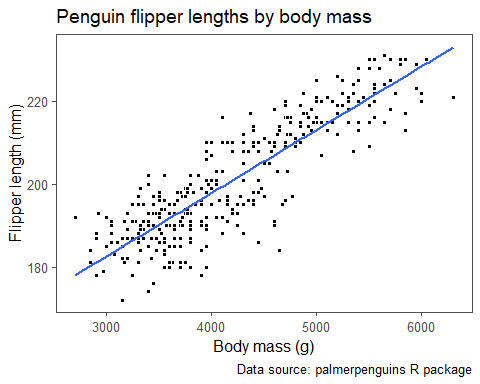
For two quantitative variables, the scatter plot is the ideal visualization. Let’s look at the relationship between the penguin flipper lengths and body masses.

*# Creating a scatter plot*  
penguins %>%   
 ggplot(aes(x = body\_mass\_g, y = flipper\_length\_mm)) +   
 geom\_point(pch = 21, color = "white", fill = "black") +  
 labs(title = "Penguin flipper lengths by body mass",  
 x = "Body mass (g)",  
 y = "Flipper length (mm)",  
 caption = "Data source: palmerpenguins R package")



We can also add a straight line of best fit to scatter plots as well.

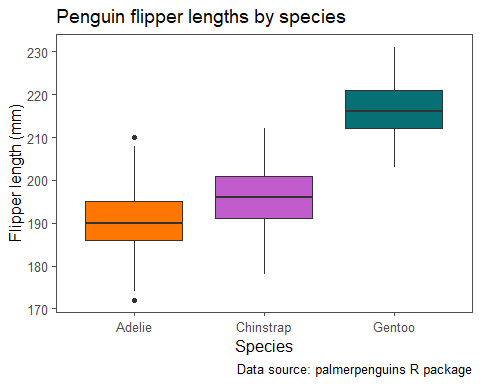
*# Creating a scatter plot with a line of best fit*  
penguins %>%   
 ggplot(aes(x = body\_mass\_g, y = flipper\_length\_mm)) +   
 geom\_point(pch = 21, color = "white", fill = "black") +  
 geom\_smooth(method = 'lm', se = FALSE) +  
 labs(title = "Penguin flipper lengths by body mass",  
 x = "Body mass (g)",  
 y = "Flipper length (mm)",  
 caption = "Data source: palmerpenguins R package")



➡️ What do we observe based on the scatter plot above?

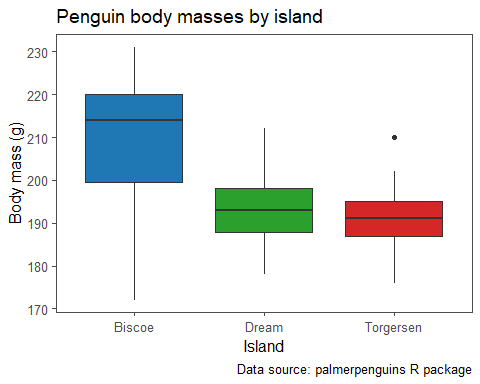
For one quantitative and one categorical variable, side-by-side box plots are a useful visualization. We can use a side-by-side boxplot to explore the penguin flipper lengths by species as below.

*# Creating side-by-side box plots*  
penguins %>%   
 ggplot(aes(x = species, y = flipper\_length\_mm, fill = species)) +   
 geom\_boxplot() +   
 scale\_fill\_manual(values = c("#ff7600", "#c05ccb", "#067075")) +  
 labs(title = "Penguin flipper lengths by species",  
 x = "Species",  
 y = "Flipper length (mm)",  
 caption = "Data source: palmerpenguins R package") +  
 theme(legend.position = "none")



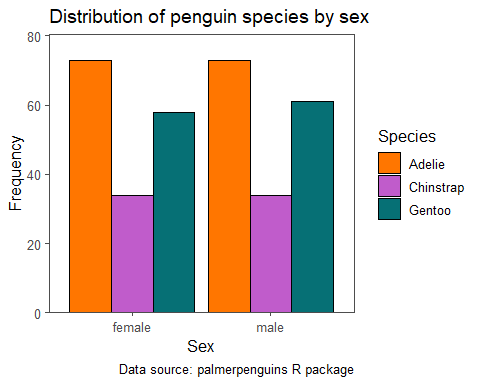
➡️ What do we observe based on the side-by-side box plots?

➡️ Modifying the previously provided code, recreate the side-by-side box plots below.

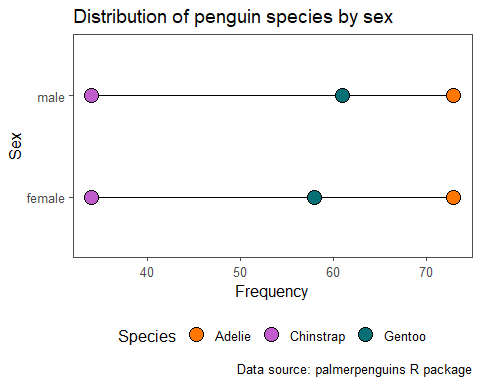


Lastly, for two categorical variables, a clustered bar chart or dumbbell chart are the more commonly used visualizations.

*# Creating a clustered bar chart*  
penguins %>% dplyr::count(species, sex, .drop = FALSE) %>%   
 dplyr::filter(!is.na(species), !is.na(sex)) %>%   
 mutate(sex = fct\_reorder(sex, n)) %>%   
 ggplot(aes(x = sex, y = n,  
 fill = species)) +   
 geom\_col(position="dodge", color = "black") +  
 scale\_fill\_manual(values = c("#ff7600", "#c05ccb", "#067075")) +  
 scale\_y\_continuous(expand = expansion(mult = c(0, 0.10))) +  
 labs(title = "Distribution of penguin species by sex",  
 x = "Sex",  
 y = "Frequency",  
 caption = "Data source: palmerpenguins R package",  
 fill = "Species")



*# Creating dumbbell chart*  
penguins %>% dplyr::count(species, sex, .drop = FALSE) %>%   
 dplyr::filter(!is.na(species), !is.na(sex)) %>%   
 dplyr::mutate(species\_sex = str\_c(species, "\_", sex)) %>%   
 ggplot(aes(x = n, y = sex,  
 color = species, fill = species)) +   
 geom\_line(aes(group = sex), color = "black") +  
 geom\_point(pch = 21, color = "black", size = 5) +  
 scale\_fill\_manual(values = c("#ff7600", "#c05ccb", "#067075")) +  
 labs(title = "Distribution of penguin species by sex",  
 x = "Frequency",  
 y = "Sex",  
 caption = "Data source: palmerpenguins R package",  
 fill = "Species") +  
 theme(legend.position = "bottom")



➡️ What can we say about the penguin sex and species based on the clustered bar chart and dumbbell chart?

### Multivariate analyses

Descriptive statistics could also be obtained jointly for three or more variables, although this is less commonly done. For example, we could obtain counts for the cross-section of the species, island, and sex variables.

*# Creating frequency table*  
speciesIslandSexCounts <- penguins %>%   
 dplyr::count(species, island, sex, .drop = FALSE)  
  
*# Printing frequency table*  
speciesIslandSexCounts %>%   
 make\_flex(caption = "Number of penguins by island, species, and sex.")

*Table 8: Number of penguins by island, species, and sex.*

| species | island | sex | n |
| --- | --- | --- | --- |
| Adelie | Biscoe | female | 22 |
| Adelie | Biscoe | male | 22 |
| Adelie | Dream | female | 27 |
| Adelie | Dream | male | 28 |
| Adelie | Dream | NA | 1 |
| Adelie | Torgersen | female | 24 |
| Adelie | Torgersen | male | 23 |
| Adelie | Torgersen | NA | 5 |
| Chinstrap | Dream | female | 34 |
| Chinstrap | Dream | male | 34 |
| Gentoo | Biscoe | female | 58 |
| Gentoo | Biscoe | male | 61 |
| Gentoo | Biscoe | NA | 5 |

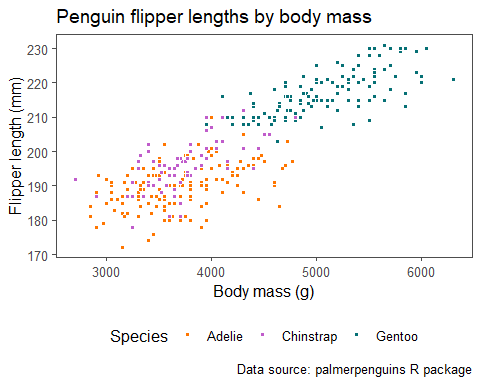
With this many variables, however, it is often better to use data visualizations to communicate aspects of the data.

### Multivariate visualizations

If we want to jointly visualize three variables, we can use a scatter plot coloring the points based on a third variable.

Let’s look at the relationship between the penguin flipper lengths and body masses while coloring the points based on the penguin species.

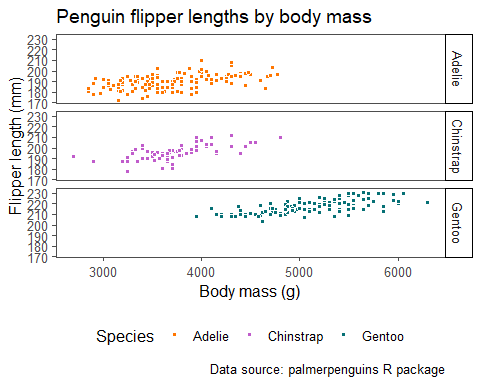
*# Creating a scatter plot*  
penguins %>%   
 ggplot(aes(x = body\_mass\_g, y = flipper\_length\_mm, fill = species)) +   
 geom\_point(pch = 21, color = "white") +  
 scale\_fill\_manual(values = c("#ff7600", "#c05ccb", "#067075")) +  
 labs(title = "Penguin flipper lengths by body mass",  
 x = "Body mass (g)",  
 y = "Flipper length (mm)",  
 fill = "Species",  
 caption = "Data source: palmerpenguins R package") +  
 theme(legend.position = "bottom")



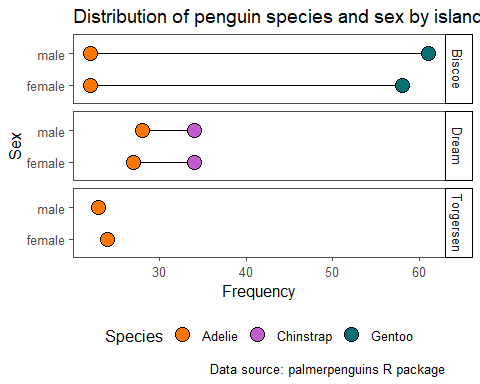
➡️ What can we say about the relationship between the penguin flipper lengths (mm) and body masses (g) for each penguin species based on the scatter plot?

An alternative way to incorporate the information of an additional categorical variable in a plot is to use *faceting*, breaking up the plot into multiple subplots. Below we showcase the relationship between the penguin flipper lengths (mm) and body masses (g) for each penguin species using faceting.

*# Creating a faceted scatter plot*  
penguins %>%   
 ggplot(aes(x = body\_mass\_g, y = flipper\_length\_mm, fill = species)) +   
 geom\_point(pch = 21, color = "white") +  
 scale\_fill\_manual(values = c("#ff7600", "#c05ccb", "#067075")) +  
 facet\_grid(species ~ .) +  
 labs(title = "Penguin flipper lengths by body mass",  
 x = "Body mass (g)",  
 y = "Flipper length (mm)",  
 fill = "Species",  
 caption = "Data source: palmerpenguins R package") +  
 theme(legend.position = "bottom",  
 strip.background.y = element\_rect(linetype = "solid", color = "black"))



For three categorical variables, we could create a faceted dumbbell chart to visualize the distribution of penguins in terms of their species, sex, and which island they were found on (Biscoe, Dream, or Torgersen).



➡️ What are some main takeaways from the faceted dumbbell chart?