Week2\_classActivity

Mallikarjuna Rao Mannem

2023-09-19

## Installed packages

# Load necessary packages  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.3 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.3 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(ggthemes)  
library(flextable)

##   
## Attaching package: 'flextable'  
##   
## The following object is masked from 'package:purrr':  
##   
## compose

library(corrr)

# 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")

# Importing Penguins data

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

## Rows: 344 Columns: 8  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (3): species, island, sex  
## dbl (5): bill\_length\_mm, bill\_depth\_mm, flipper\_length\_mm, body\_mass\_g, year  
##   
## 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.

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

### 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 |

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

The largest was 231mm and the smallest was 172mm.

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

M = 197mm, so 50% of penguins in the data set had a flipper length 197mm or longer.

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

Since s = 14.06, the sample variance is s2= 197.6836mm^2

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

The range is more sensitive to outliers - the IQR is more robust to outliers than the range.

### 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 |

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

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?**

Adelie with n=152.

**Which island had the least penguins measured on it?**

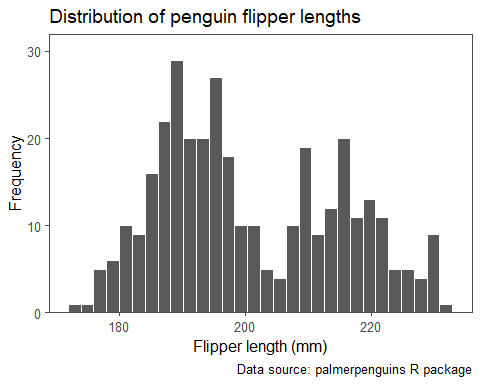
Torgersen with n=52.

### Recreate the histogram and a box plot for the penguin flipper lengths

# 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")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

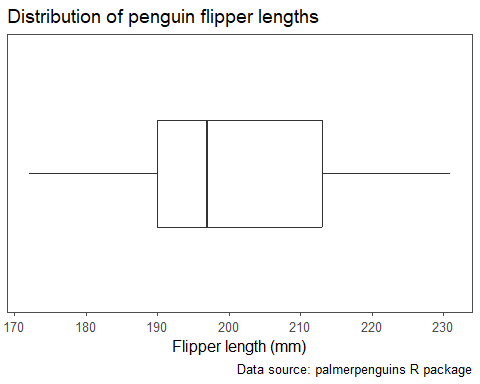
## Warning: Removed 2 rows containing non-finite values (`stat\_bin()`).



### Box Plot

# 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")

## Warning: Removed 2 rows containing non-finite values (`stat\_boxplot()`).



**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?**

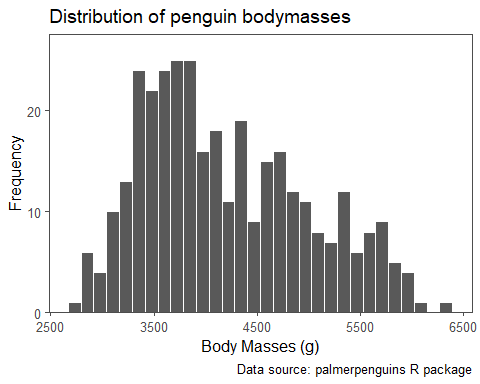
Box plot cannot say about the modal type. No outliers.

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

# Creating a histogram  
penguins %>%   
 ggplot(aes(x = body\_mass\_g)) +   
 geom\_histogram(color = "white") +  
 scale\_y\_continuous(expand = expansion(mult = c(0, 0.10))) +  
 labs(title = "Distribution of penguin bodymasses",  
 x = "Body Masses (g)",  
 y = "Frequency",  
 caption = "Data source: palmerpenguins R package")

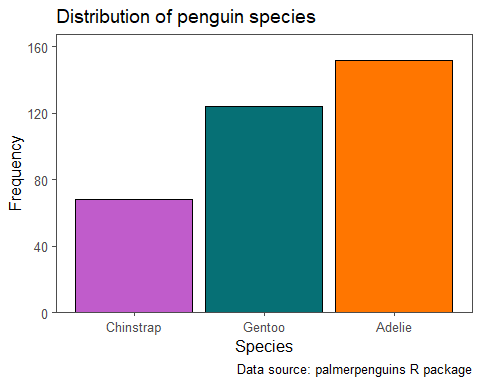
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 2 rows containing non-finite values (`stat\_bin()`).



### 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")

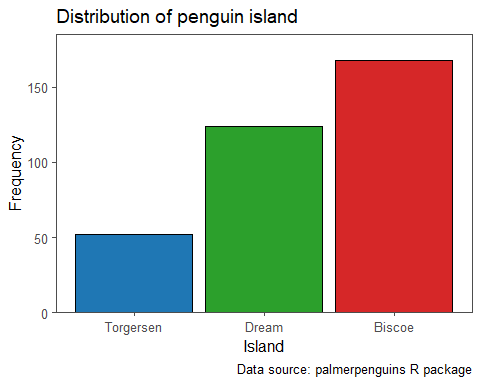


**What can we say about the penguin species based on the bar chart?**

Linear and Unimodal.

**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”).**

# Creating a bar chart  
  
penguins %>% dplyr::count(island, .drop = FALSE) %>%   
 mutate(island = fct\_reorder(island, n)) %>%   
 ggplot(aes(x = island, y = n,  
 fill = island)) +   
 geom\_col(color = "black") +  
 scale\_fill\_manual(values = c("#1F77B4", "#2CA02C", "#D62728")) +  
 scale\_y\_continuous(expand = expansion(mult = c(0, 0.10))) +  
 labs(title = "Distribution of penguin island",  
 x = "Island",  
 y = "Frequency",  
 caption = "Data source: palmerpenguins R package") +  
 theme(legend.position = "none")



## Bivariate descriptive statistics

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

# Calculating correlations  
corTable <- penguins %>%   
 corrr::correlate(diagonal = 1)

## Non-numeric variables removed from input: `species`, `island`, and `sex`  
## Correlation computed with  
## • Method: 'pearson'  
## • Missing treated using: 'pairwise.complete.obs'

# 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?**

flipper\_length\_mm and body\_mass\_g have the strongest correlation with r = 0.87.

**Which variables have the weakest correlation?**

body\_mass\_g and year has weakest correlation with r = 0.04.

### 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?**

Gentoo species has the largest flipper lengths.

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

Chinstrap species has the most variability in the flipper lengths.

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

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

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 |

### Frequency Table

# 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?**

Yes, Adelie species is in three different islands.

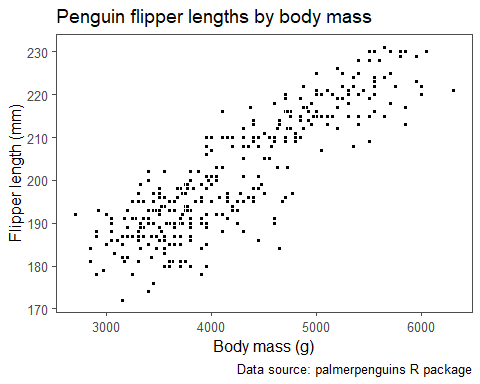
**How many penguins were found on Dream island?**

124 penguins found in Dream island.

### Bivariate visualizations

# 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")

## Warning: Removed 2 rows containing missing values (`geom\_point()`).



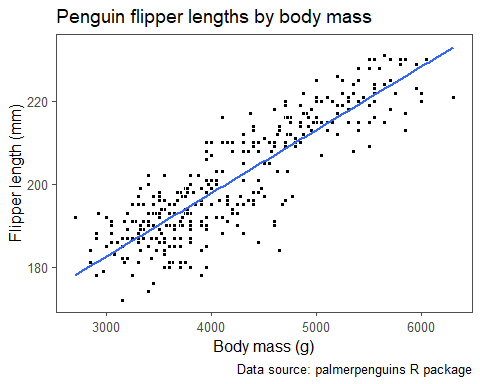
### We can also add a straight line of best fit to scatter plots as well.

# 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")

## `geom\_smooth()` using formula = 'y ~ x'

## Warning: Removed 2 rows containing non-finite values (`stat\_smooth()`).

## Warning: Removed 2 rows containing missing values (`geom\_point()`).

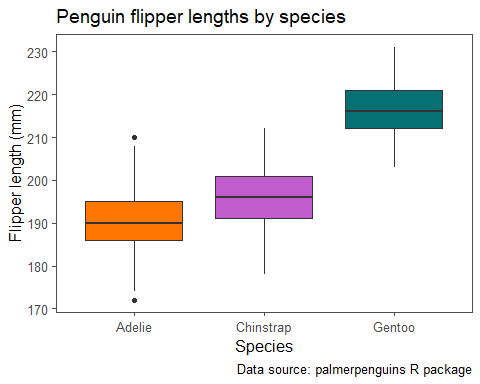
 **What do we observe based on the scatter plot above?**

Plot shows the data is linear.

### 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")

## Warning: Removed 2 rows containing non-finite values (`stat\_boxplot()`).



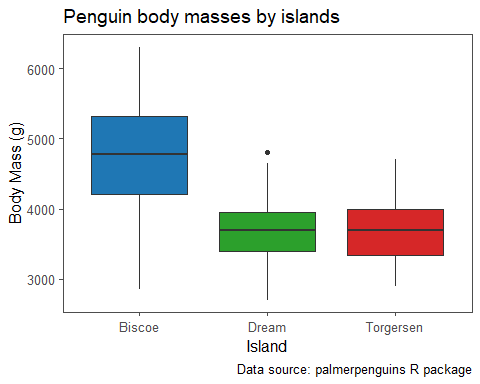
**What do we observe based on the side-by-side box plots?**

All the species have similar variability and Adelie has minimum flipper length and Gentoo has maximum flipper length. Adelie Species has outliers but not too far away.

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

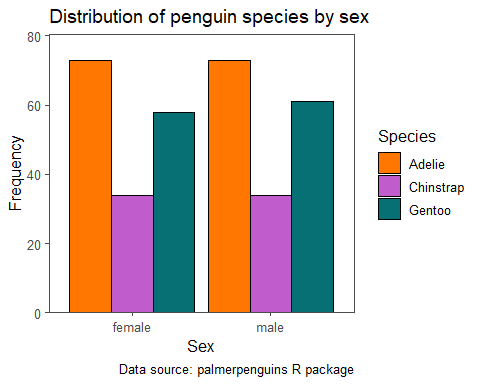
# Creating side-by-side box plots  
penguins %>%   
 ggplot(aes(x = island, y = body\_mass\_g, fill = island)) +   
 geom\_boxplot() +   
 scale\_fill\_manual(values = c("#1F77B4", "#2CA02C", "#D62728")) +  
 labs(title = "Penguin body masses by islands",  
 x = "Island",  
 y = "Body Mass (g)",  
 caption = "Data source: palmerpenguins R package") +  
 theme(legend.position = "none")

## Warning: Removed 2 rows containing non-finite values (`stat\_boxplot()`).



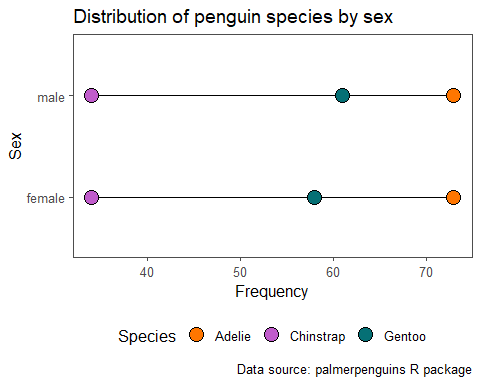
### Lastly, for two categorical variables, a clustered bar chart or dumbbell chart are the more commonly used visualizations.

# 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")



### Dumbbell Chart

# 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?**

Male Gentoo Species are more in number than female Gentoo Species and the rest of species are same in number with respect to sex.

## 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.

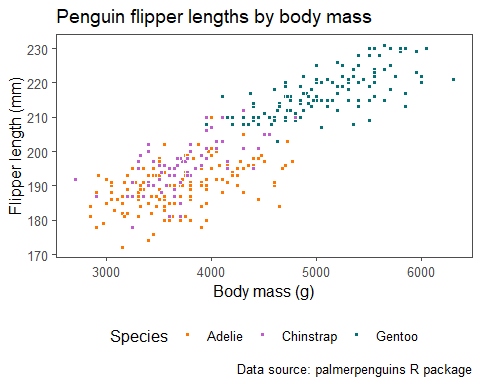
| species | island | sex | n |
| --- | --- | --- | --- |
| Adelie | Biscoe | female | 22 |
| Adelie | Biscoe | male | 22 |
| Adelie | Dream | female | 27 |
| Adelie | Dream | male | 28 |
| Adelie | Dream | NA | 1 |
| Adelie | Torgersen | female | 24 |
| Adelie | Torgersen | male | 23 |
| Adelie | Torgersen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

## 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")

## Warning: Removed 2 rows containing missing values (`geom\_point()`).



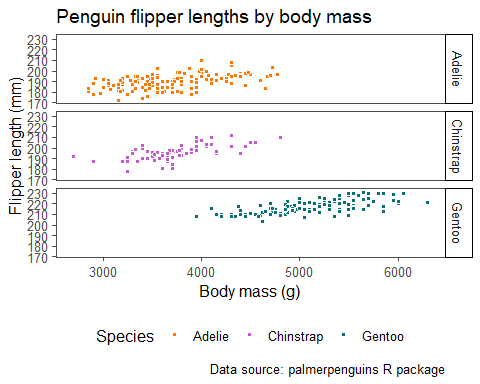
**What can we say about the relationship between the penguin flipper lengths (mm) and body masses (g) for each penguin species based on the scatter plot?**

If we consider the linearity of each species, their slopes are different.

### 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"))

## Warning: Removed 2 rows containing missing values (`geom\_point()`).



### Faceted Dumbell chart

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

Gentoo is large in number and is found only in Biscoe island.

Chinstrap is low in number and is found only in Dream island.

Adelie is lowest in number in each island and is found in all islands.