# Code-a-long-2.2

# Box plots and error bars

## Goals

1. Students will be able to create box plots from data sets.
2. Students will be able to add error bars to bar plots.
3. Students will be able to interpret a box plots and the purpose of error bars in bar plots.

For these examples, we're going to use the Palmer penguins data set.

```
library(tidyverse)
```

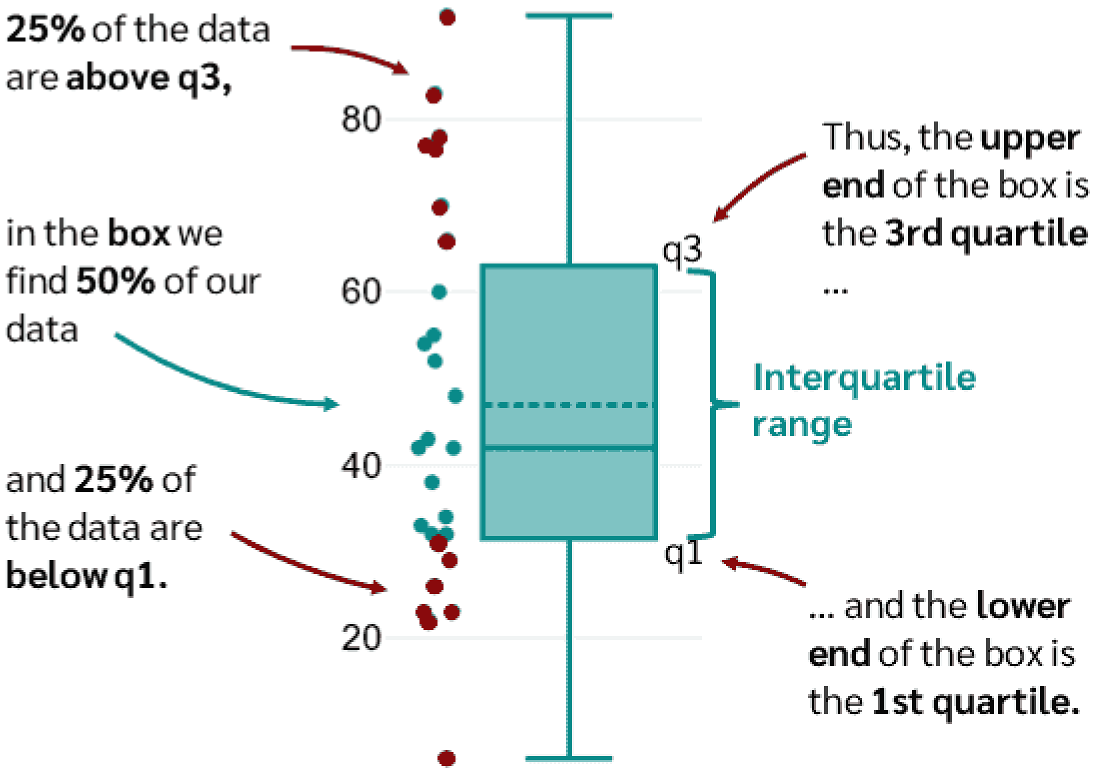
```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✔ dplyr     1.1.2     ✔ readr     2.1.4
✔ forcats   1.0.0     ✔ stringr   1.5.0
✔ ggplot2   3.4.2     ✔ tibble    3.2.1
✔ lubridate 1.9.2     ✔ tidyr     1.3.0
✔ purrr     1.0.2
── Conflicts ────────────────────────────────────────── 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(palmerpenguins)

penguins<-palmerpenguins::penguins

penguins
```

```
# A tibble: 344 × 8
   species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
   <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
 1 Adelie  Torgersen           39.1          18.7               181        3750
 2 Adelie  Torgersen           39.5          17.4               186        3800
 3 Adelie  Torgersen           40.3          18                 195        3250
 4 Adelie  Torgersen           NA            NA                  NA          NA
 5 Adelie  Torgersen           36.7          19.3               193        3450
 6 Adelie  Torgersen           39.3          20.6               190        3650
 7 Adelie  Torgersen           38.9          17.8               181        3625
 8 Adelie  Torgersen           39.2          19.6               195        4675
 9 Adelie  Torgersen           34.1          18.1               193        3475
10 Adelie  Torgersen           42            20.2               190        4250
# ℹ 334 more rows
# ℹ 2 more variables: sex <fct>, year <int>
```

## Box plots with error bars

This image illustrates a box plot, and how to interpret it:


**25%** of the data are **above q3,**
in the **box** we find **50%** of our data
and **25%** of the data are **below q1.**
80
60
40
20
q3
q1
Thus, the **upper end** of the box is the **3rd quartile** ...
Interquartile range
... and the **lower end** of the box is the **1st quartile.**

*From https://datatab.net/tutorial/box-plot*

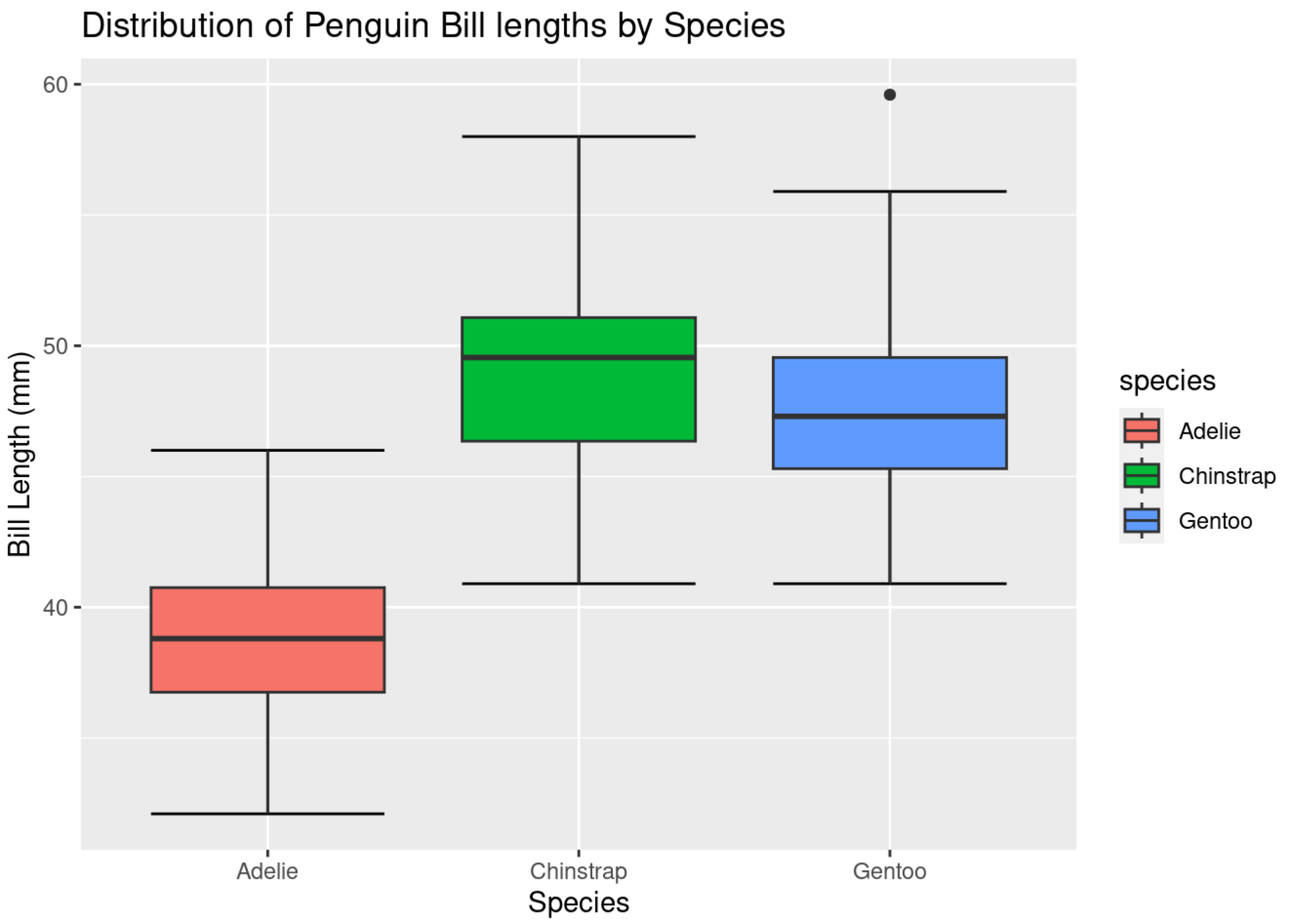
Sometimes you'll see points that extend beyond the whiskers. These points are considered outliers, as they significantly deviate from the rest of the data in the data set.

Let's see how the measurements of bill lengths compare among penguin species using a box plot.

Strategy: Use `geom_boxplot`, and in the `aes` function, set x equal to a categorical column, which will automatically group them. For the horizontal lines (error bars), add stat_boxplot(geom = "errorbar").

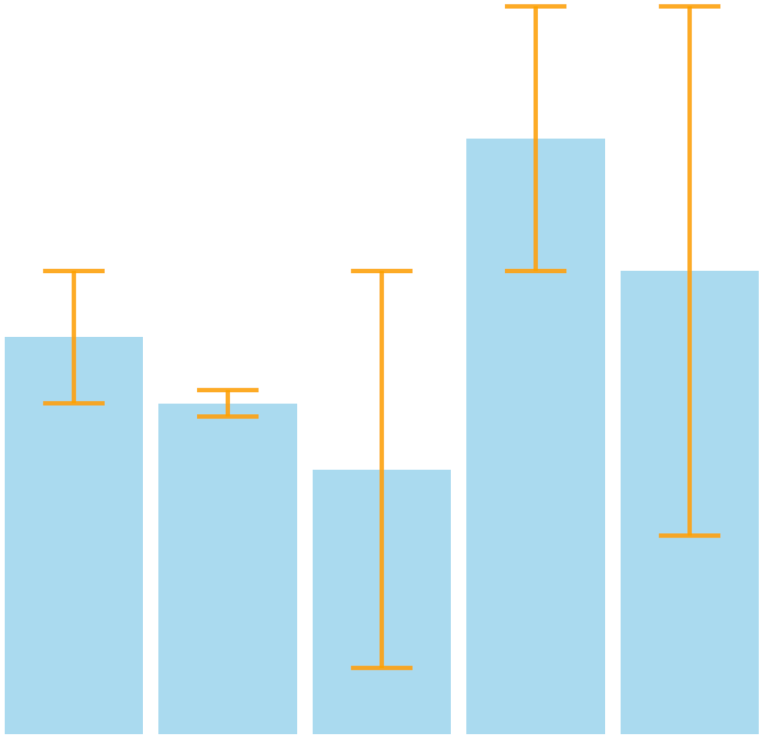
```
ggplot(data=penguins, mapping=aes(x=species, y=bill_length_mm, fill=species))+
  stat_boxplot(geom="errorbar")+
  geom_boxplot()+
  labs(title="Distribution of Penguin Bill lengths by Species",
       x="Species", y="Bill Length (mm)")
```

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



## Bar plots with error bars

“Error bars” can be included in bar plots to display the variability in the data. The term doesn’t necessarily imply that the data is erroneous, but is important for showing the range of measurement. A common approach is to use the mean +/- one standard deviation to calculate the error bars.



from https://r-graph-gallery.com/4-barplot-with-error-bar.html

Typically we're interested in determining whether one population differs from another on some trait. If error bars are large and/or overlapping, we're less likely to determine that the samples are different. Smaller/non-overlapping error bars indicate that there may be a true difference between bars.

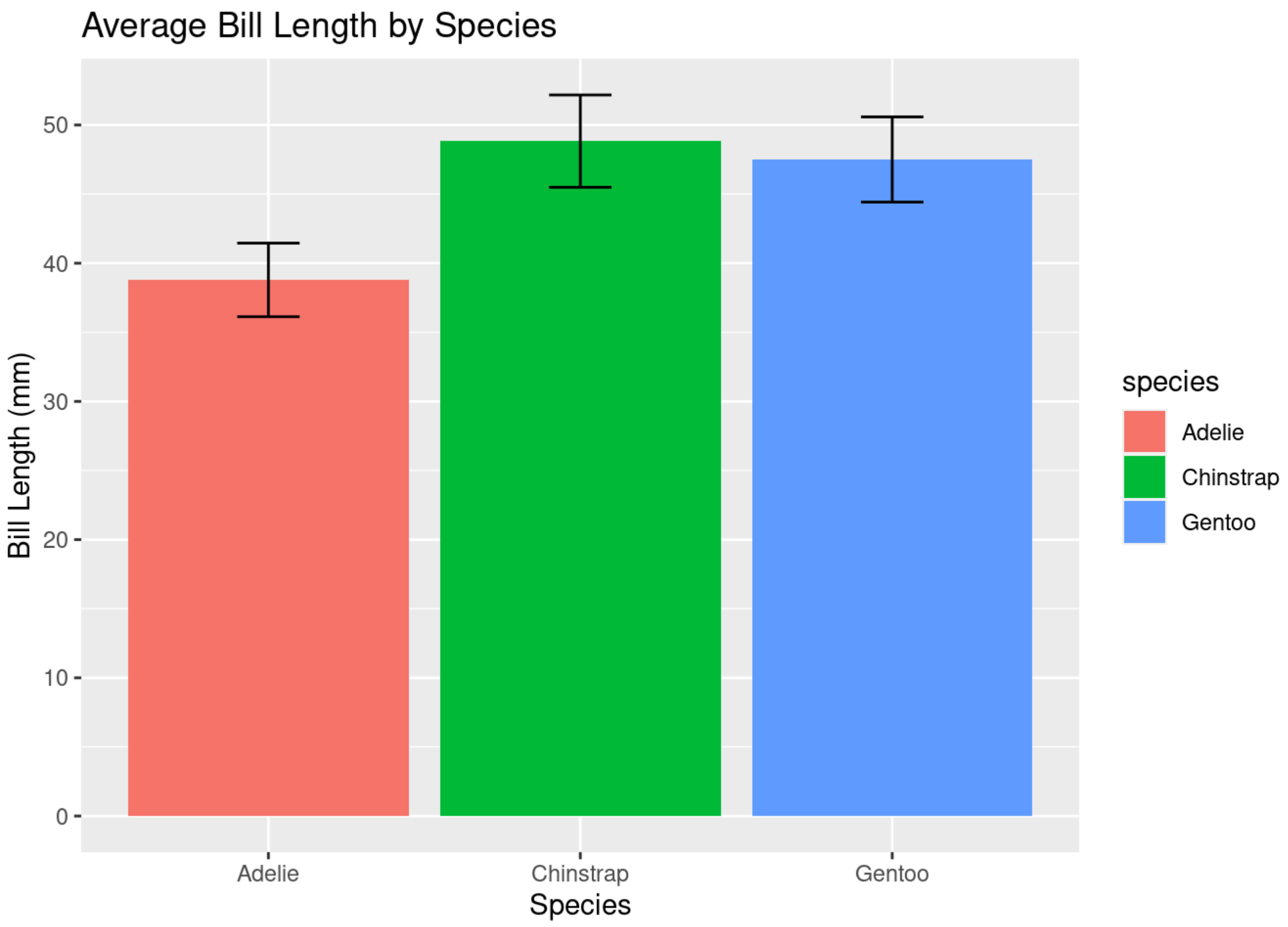
Let's do the same comparison of bill length by species, now using a bar plot with error bars.

Strategy: use `group_by`/`summarize` to calculate group mean and standard deviation. Add `geom_errorbar`, in which the top top bar is the mean + standard deviation, and the lower bar is the mean - standard deviation:

```
groupedBillLength<- penguins |>
  group_by(species) |>
  summarize(avgBillLength=mean(bill_length_mm, na.rm=TRUE),
            sdBillLength=sd(bill_length_mm, na.rm=TRUE))

groupedBillLength
```

```
# A tibble: 3 × 3
  species   avgBillLength sdBillLength
  <fct>             <dbl>        <dbl>
1 Adelie             38.8         2.66
2 Chinstrap          48.8         3.34
3 Gentoo             47.5         3.08
```

```
ggplot(data=groupedBillLength, mapping=aes(x=species,
                                           y=avgBillLength,
                                           fill=species))+
  geom_bar(stat="identity")+
  geom_errorbar(mapping=aes(ymin=avgBillLength-sdBillLength,
                            ymax=avgBillLength+sdBillLength), width=0.2)+
  labs(title="Average Bill Length by Species",
       x="Species",
       y="Bill Length (mm)")
```

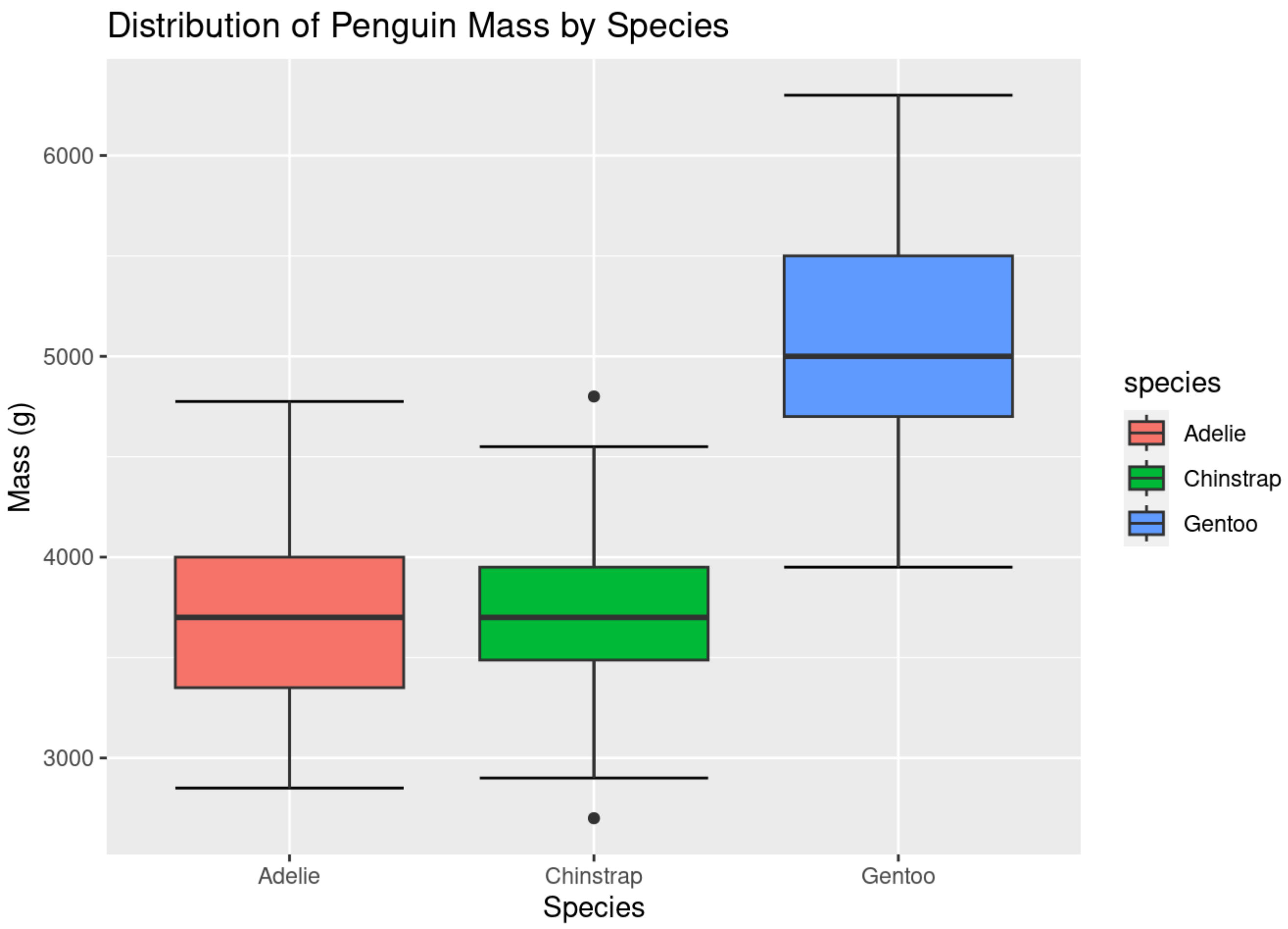


## Practice

Create a box plot comparing penguin mass among species. Include error bars, a legend, and labels.

```
# create box plot below

ggplot(data=penguins, mapping=aes(x=species, y=body_mass_g, fill=species))+
  stat_boxplot(geom="errorbar")+
  geom_boxplot()+
  labs(title="Distribution of Penguin Mass by Species",
       x="Species", y="Mass (g)")
```

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

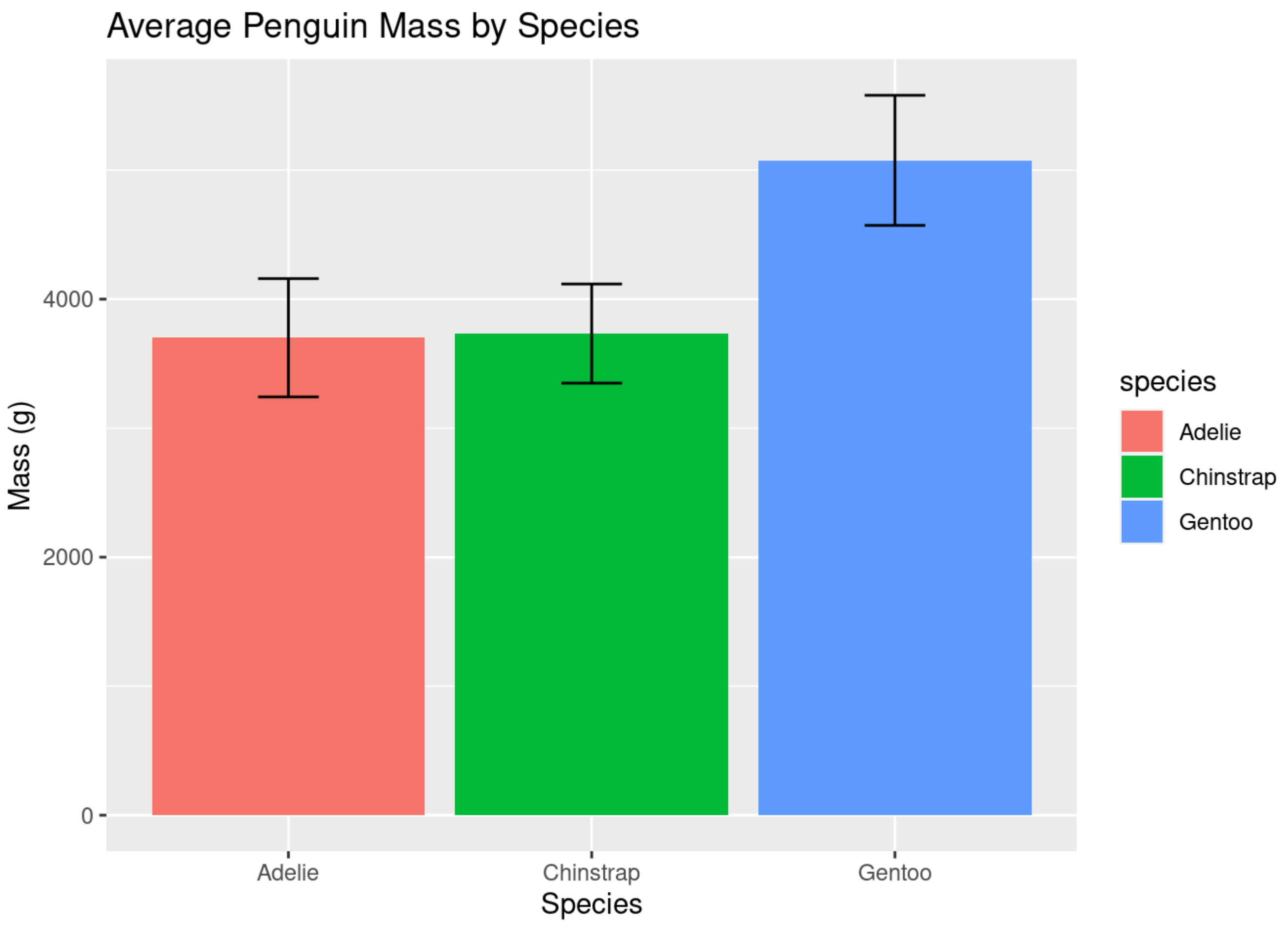


Create a bar plot comparing penguin mass among species. Include error bars (+/- 1 sd), a legend, and labels.

```
# create bar plot below



groupedMass<- penguins %>%
  group_by(species) %>%
  summarize(avgMass=mean(body_mass_g, na.rm=TRUE),
            sdMass=sd(body_mass_g, na.rm=TRUE))

groupedMass
```

```
# A tibble: 3 × 3
  species   avgMass sdMass
  <fct>       <dbl>  <dbl>
1 Adelie      3701.   459.
2 Chinstrap   3733.   384.
3 Gentoo      5076.   504.
```

```
ggplot(data=groupedMass, mapping=aes(x=species, y=avgMass, fill=species))+
  geom_bar(stat="identity")+
  geom_errorbar(mapping=aes(ymin=avgMass-sdMass,
                            ymax=avgMass+sdMass), width=0.2)+
  labs(title="Average Penguin Mass by Species", x="Species", y="Mass (g)")
```



Next: code-a-long 2.3