Day 6 - intro to ggplot2

Introduction

The purpose of this activity is to use some of the `tidyverse` skills we developed yesterday to create nice graphics via `ggplot2`. You will likely need to consult me, R help files, Google, or ChatGPT to fully complete this activity (*that is how coding is generally learned*).

The structure of `ggplot2`

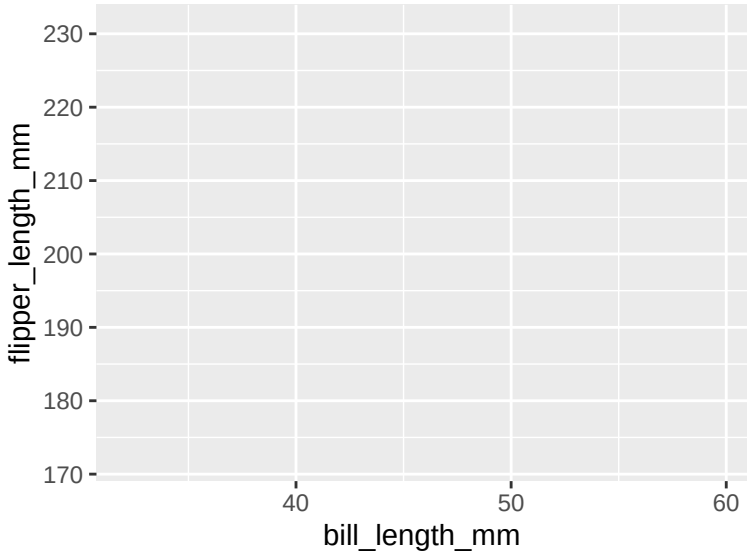
> Note
>
> `ggplot2` functions by layering calls to various functions on top of one another to construct a complete plot. Generally, the steps are as follows:

1) **Create an empty canvas**

```
penguins %>%
  ggplot()
```

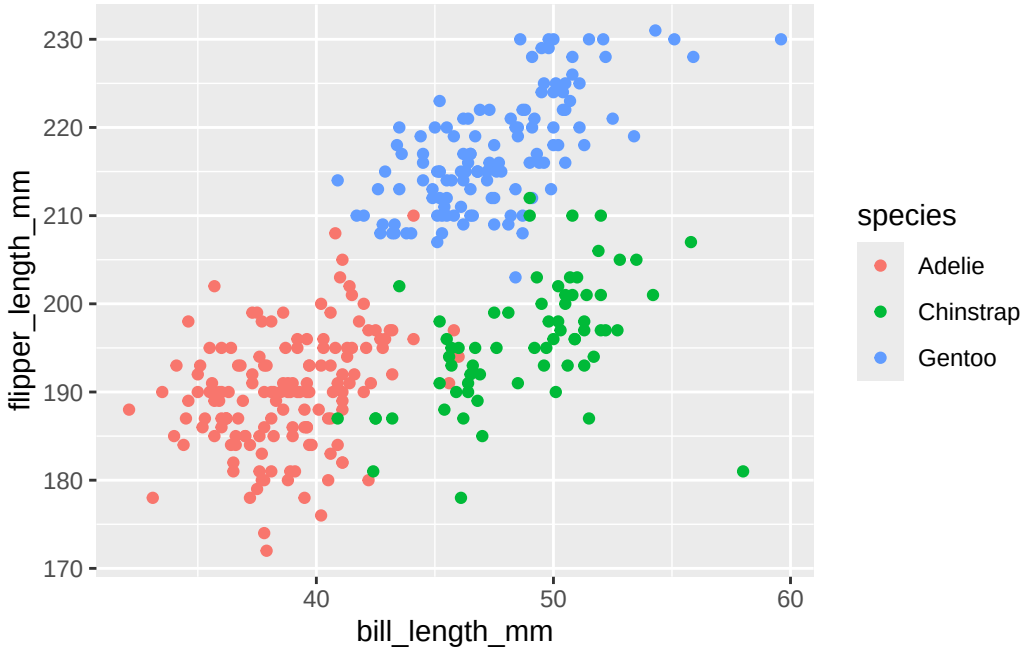
2) **Add plotting aesthetics**

```
penguins %>%
  ggplot(aes(x = bill_length_mm, y = flipper_length_mm, col = species))
```



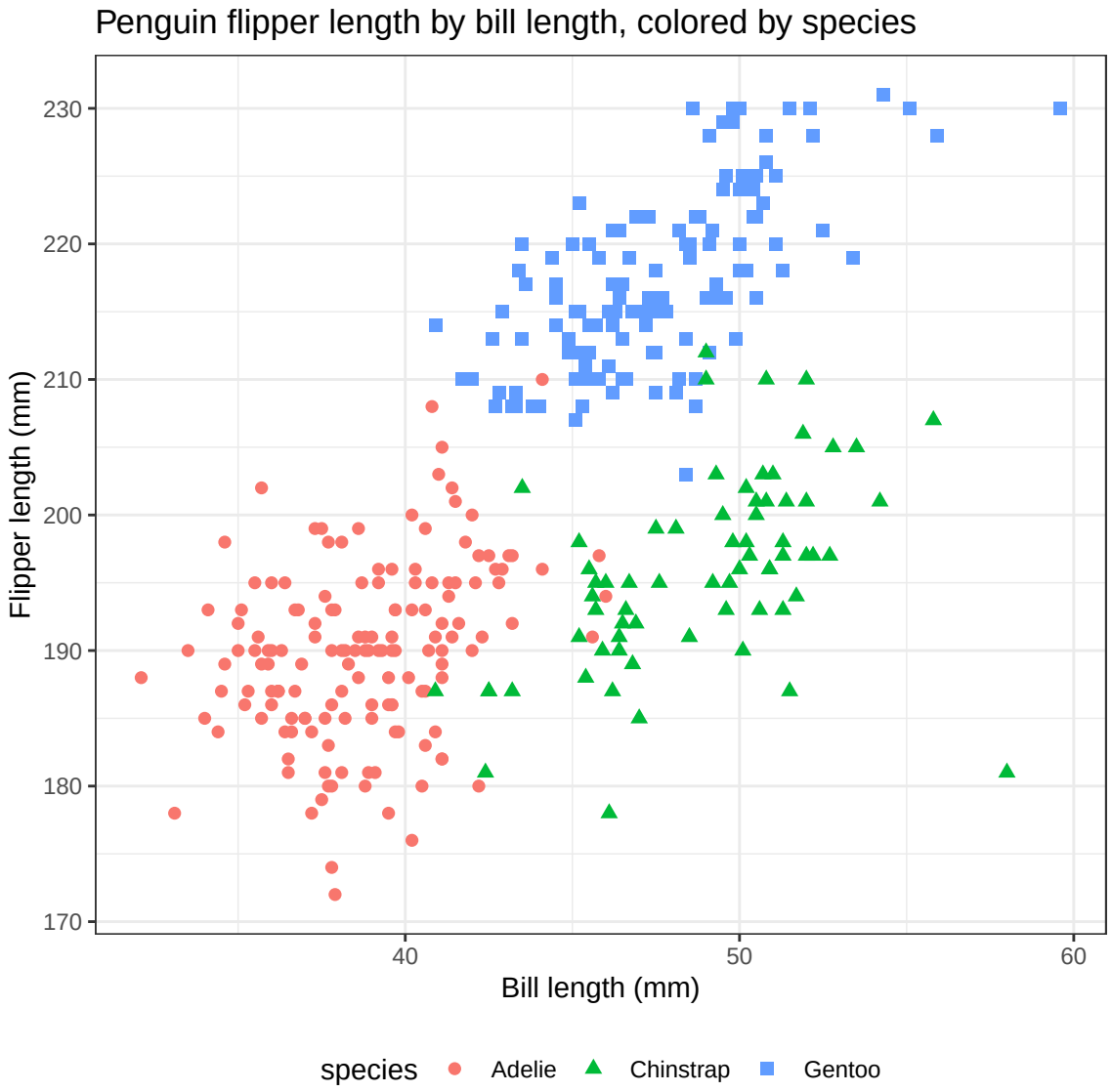
3) **Add geometries**

```
penguins %>%
  ggplot(aes(x = bill_length_mm, y = flipper_length_mm, col = species)) +
  geom_point()
```



4) **Formatting and options**

```
penguins %>%
  ggplot(aes(x = bill_length_mm, y = flipper_length_mm, col = species)) +
  geom_point(aes(shape = species), size = 2) +
  theme_bw() +
  labs(
    title = "Penguin flipper length by bill length, colored by species",
    x = "Bill length (mm)",
    y = "Flipper length (mm)"
  ) +
  theme(legend.position = "bottom")
```



Activity

Quantitative responses

Research has shown that, for most mammals, as body mass increases, brain size tends to increase as well. This relationship is referred to as the "raw brain-to-body mass ratio." This ratio, while not wholly accurate, is used as a proxy to compare intelligence across species, as it is thought that larger brains are more capable of complex cognitive tasks. However, this relationship has only been examined in mammals. Researchers are interested in whether this brain-to-body mass relationship holds for avian species as well. Using a combination of methods of data collection (existing data combined with specimens from the London Museum of Natural History), researchers examined a total of 61 species of woodpeckers. The following variables were collected:

- The `Species` name
- `SocialCategory`: the type of general social structure typical for that species (solitary, pair-living, group-living)
- `BodyWeight`: the mass of the species (in grams)
- `Volume`: the volume of the brain of the species (in mm^3)

> **Note**
>
> The article[1] is vague about how the weight/volume measurements were collected. It is thought that the researchers averaged the measurements from two available specimens of each species when possible, otherwise the average measurements from existing data were used.

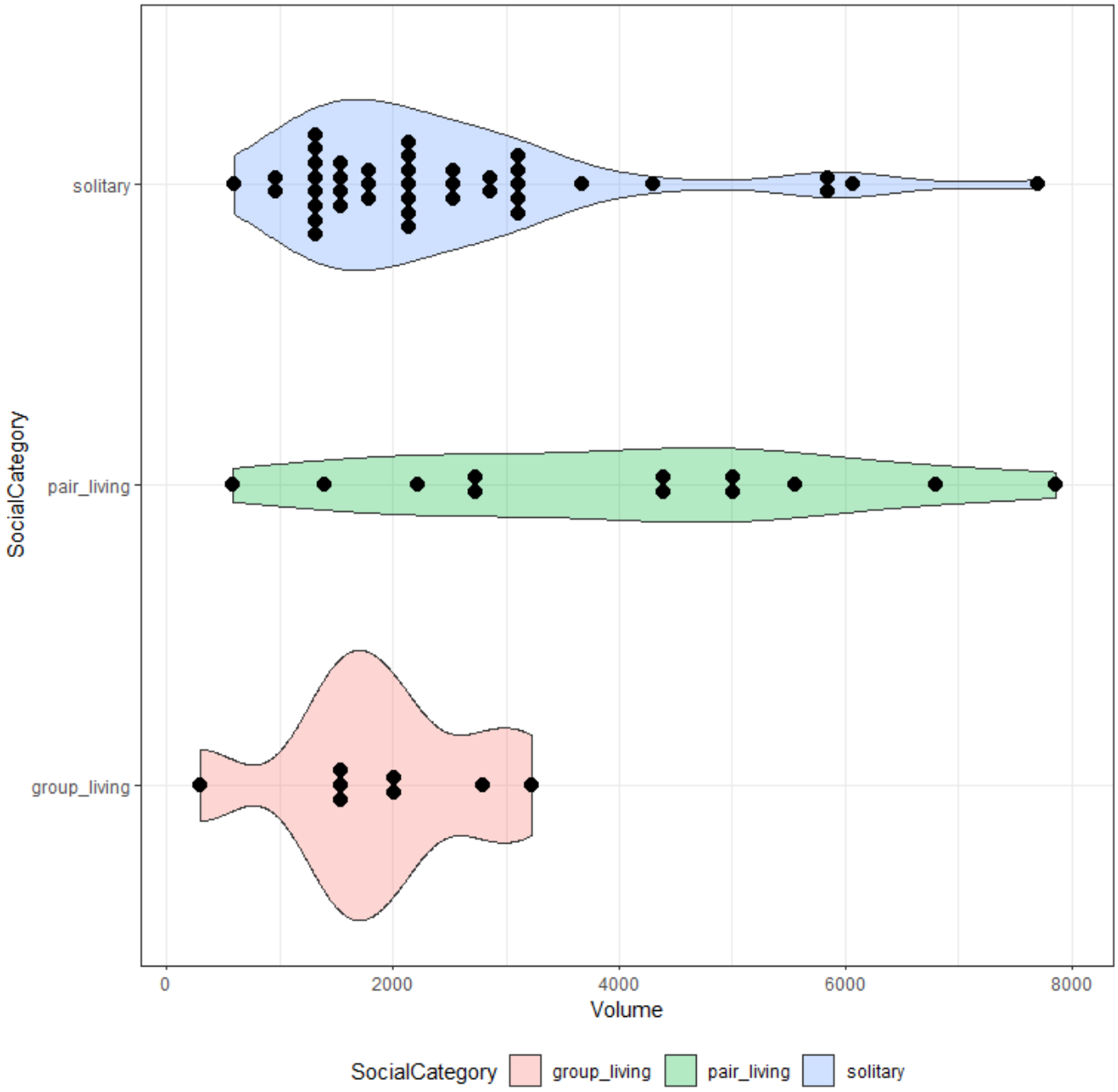
```
woodpecker <- read_csv("woodpeckers.csv")
```

1. Create a histogram of the `BodyWeight` variable and describe the distribution in terms of shape, center, and spread. (hint: try `?geom_histogram`)

2. Change the `fill` and `color` of the histogram to colors of your choosing. What do these aesthetics do?

3. Recreate the following figure (this link may be helpful https://r-charts.com/distribution/violin-plot-points-ggplot2/) and describe each distribution in terms of shape, center, and spread.

[1] See more here.



4. Researchers are interested in determining whether there is an association between body weight and brain volume for woodpeckers. Create the appropriate plot for this research question and provide a meaningful title and labels. (Hint: think about what should go on the x and y axes!)
5. Recreate the previous plot, changing the color of the points to one of your choosing.
6. Based on this visual, does there appear to be an association between brain volume and body weight? Why or why not?

7. Suppose researchers are interested in determining whether the nature of the aforementioned association differs between social categories. Create a visual that address this research question. Comment on what the visual suggests about the new research question.

8. When conducting statistical analyses, it is often useful to log transform variables to obtain nice linear relationships. Log both the explanatory and response variables and recreate the visual, adding a smoother (hint: `?geom_smooth`). What does the updated figure suggest about the research question?

Categorical responses

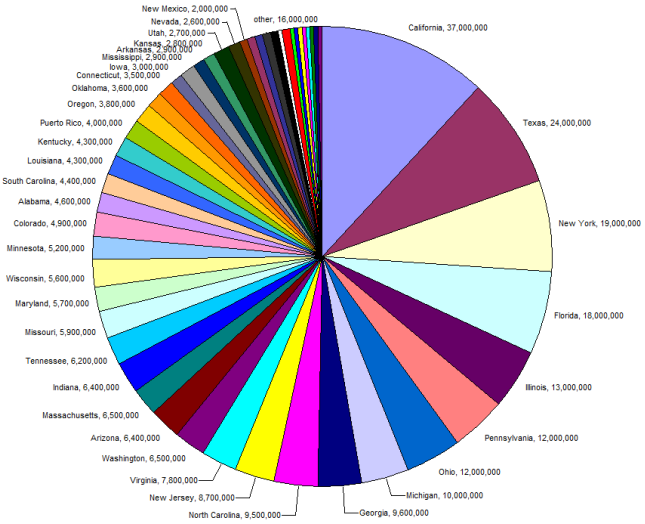
To explore categorical variables, we will use the `Titanic` data. See `?Titanic` for more information.

```
# convert high dimensional array to data frame for plotting
titanic <- Titanic %>%
  as_tibble() %>%
  uncount(n)
```

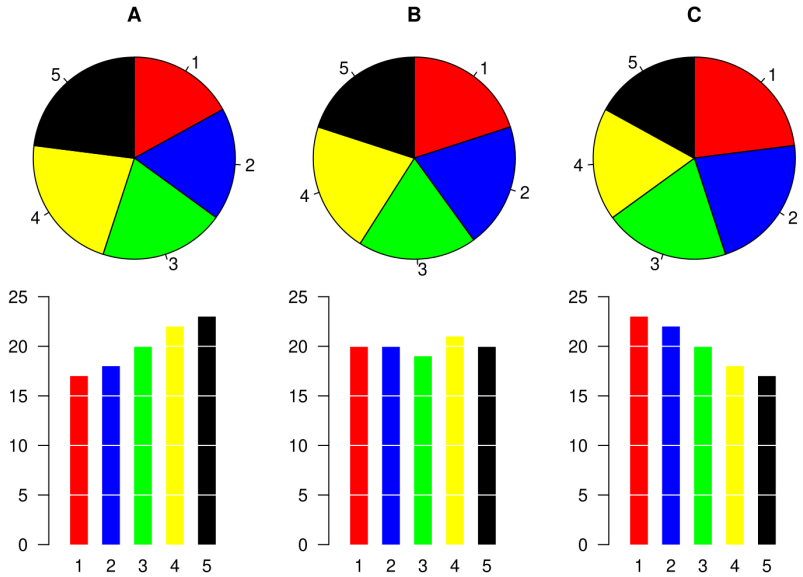
1. Use the `Titanic` data to visualize how the probability of survival differs by `Class`. For this question, you should create three bar plots - **stacked**, **dodged**, and **standardized** versions. Comment on the different information contained in each plot.

A note about pie charts

> **Note**
>
> Pie charts, while potentially useful for visualizing categorical variables with few levels, should not be used to show comparisons or trends.



Source: Wikipedia

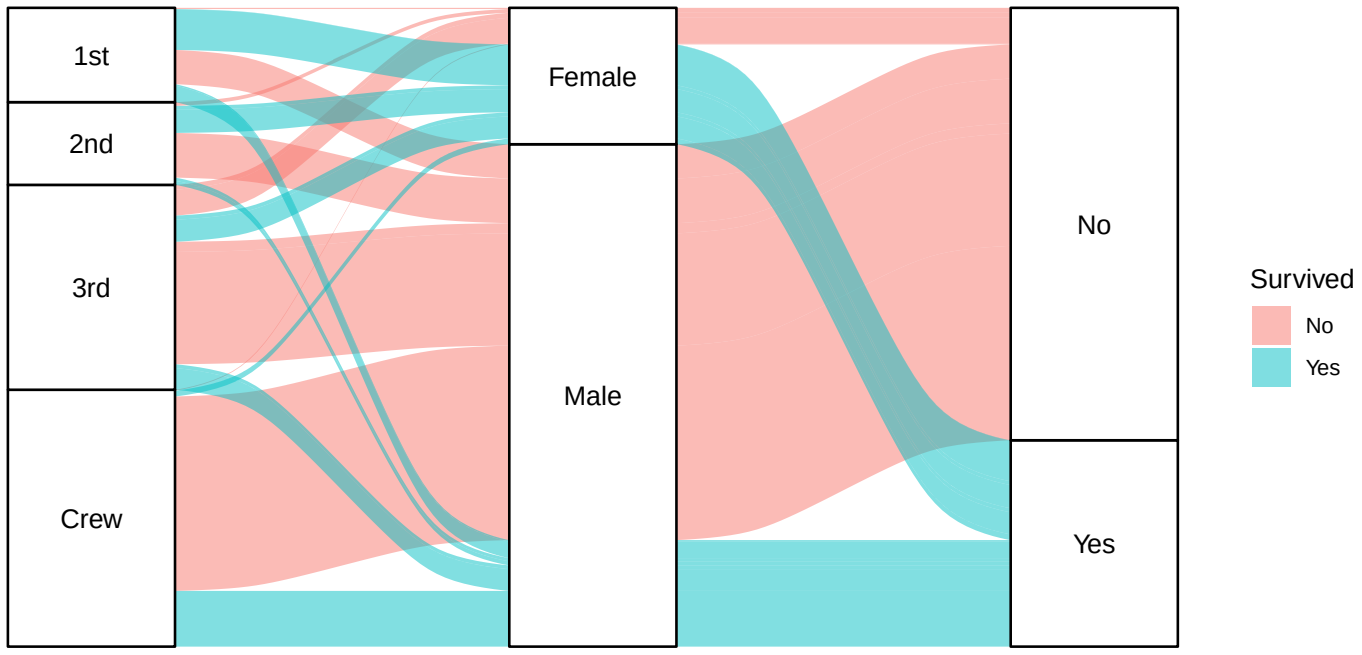


Source: Wikipedia

Other fun graphics

> **Note**
>
> An alluvial diagram[2] is useful for visualizing how categorical outcomes relate to categorical predictors.

```
# use the Titanic survival data
## see ?Titanic
library(ggalluvial)
Titanic %>%
  as_tibble %>%
  ggplot(
    aes(axis1 = Class, axis2 = Sex, axis4 = Survived, y = n)
  ) +
  geom_alluvium(aes(fill = Survived)) +
  geom_stratum() +
  geom_text(
    stat = "stratum",
    aes(label = after_stat(stratum))
  ) +
  theme_void()
```



[2] See more here.