BIOS 42411: Lab 2

Kevin Buck

1/25/2023

In class

Getting started

Today we are going to create a variety of plots that display and summarize both continuous and categorical data. We will be using the package **ggplot2** which is a part of the **tidyverse**: a collection of R packages that uses a standardized syntax to ease the transition of moving from loading data, to organizing/summarizing data, to creating plots and even analyzing data. If you have not yet installed the package **tidyverse** please do so. Installing **tidyverse** will automatically install **ggplot2** (don't forget to load the libraries!). We'll use "tidyverse" all semester. It's a nice way to organize and graph data (see tidyverse.org).

Load the data set ("penguins") by installing the package **palmerpenguins** and loading the library.

```
# Load libraries
library(tidyverse)
library(palmerpenguins)
```

About the data

Penguins are awesome animals. We're going to learn about their biology by plotting measurments of their bills. The data you are working with today were collected and made available by Dr. Kristen Gorman via the Palmer Station Antarctica Long-Term Ecological Research Station (LTER). These data were originally published in:

Gorman KB, Williams TD, Fraser WR (2014). Ecological sexual dimorphism and environmental variability within a community of Antarctic penguins (genus Pygoscelis). PLoS ONE 9(3):e90081. https://doi.org/10.1371/journal.pone.0090081

It's always a good idea to take a look at your data to get an idea of what the column names are, how many observations you have, if you have missing data, etc. You can do this using the **str()** function (str is short for "structure"). Test this out below. To check what your argument should be, check the help file by typing either **help(str)** or by pressing F1 on **Windows**/Ctrl+Option+F1 on a **Mac** while your cursor is in the function. There are often examples at the bottom of the help files.

```
str(penguins)
```

```
## tibble [344 x 8] (S3: tbl_df/tbl/data.frame)
##  $ species          : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ island           : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ bill_length_mm   : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
##  $ bill_depth_mm    : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
##  $ flipper_length_mm: int [1:344] 181 186 195 NA 193 190 181 195 193 190 ...
##  $ body_mass_g      : int [1:344] 3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 ...
##  $ sex              : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2 NA NA ...
##  $ year             : int [1:344] 2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
```

That's a lot of output! As always, when you're learning something new, pay attention to the details (that way, they'll all come more easily later): Make sure you understand each line of the str() output. If it helps, you can view the first 6 rows of data using head(penguins).

Tidyverse and ggplot basics

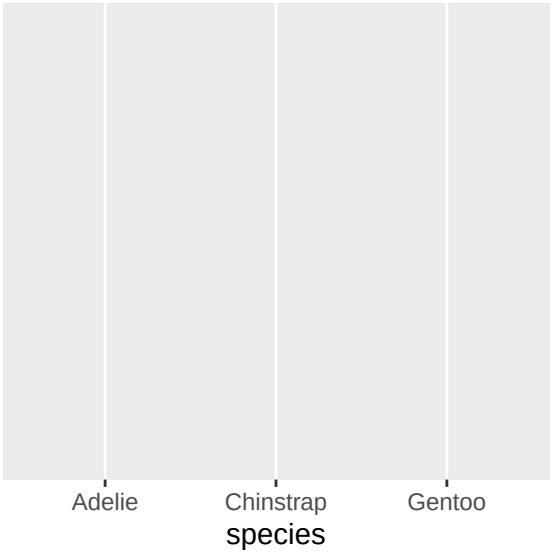
The general workflow for working with data in **ggplot2** is:

```
data %>%
  ggplot(aes(x = explanatory, y = response)) +
  geom_XX()
# Note: this code does not run and is just an example
```

The first thing to type will be the data frame, which in this lab will be **penguin**. The %>% is known in tidyverse as a **pipe** (keyboard shortcuts = *Mac*: CMD+Shift+M, *Windows*: CTRL+Shift+M). Your data object always goes before your pipe. The geom_XX determines the type of graph (*e.g.* bar plot, scatter plot, box plot). Once the data has moved through the pipe, you can access any of its columns for the arguments of your plot by simply typing, with correct capitalization, the column name. RStudio will actually prompt you with auto-complete options as you start typing the column name of interest, or you can press Tab to display the possible column names in a list that you can select from.

To create a plot, R first needs to know the **aesthetics** ("aes()"). For graphs with one variable (*i.e.* frequency distributions of categorical or continuous variables), this can just be the name of that variable. For example, if we want to get the frequency distribution of species of penguins in our dataset:

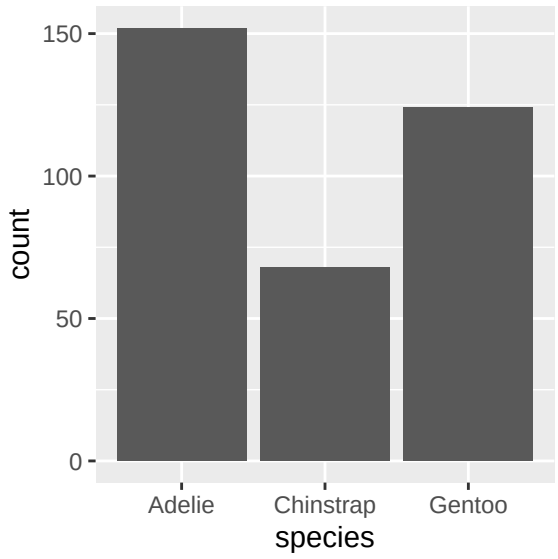
```
penguins %>%
  ggplot(aes(species))
```



You can see that all the **ggplot** function does is set up the graph; it does not actually plot the data. Plotting the data or modifying the plot in any other way is considered an "add-on" to this plot.

Now let's plot the frequency distribution of species of penguins on this graph. We will use **geom_bar()** to make a barplot because species is a categorical variable. Because **geom_bar()** is an "add-on" there is a "+" sign to add that feature to the plot. Be sure to annotate your code line by line.

```
penguins %>%  #puts data into pipe
  ggplot(aes(species)) + #creates plot with the data being species (discrete)
  geom_bar() #adds barplot of the frequency of the different species values
```



What conclusions can we draw from this plot?

This plot shows thhat Adelie penguins are most abundant, followed by Chinstrap penguins and Gentoo penguins (respectively) in this dataset.

Make your own plot

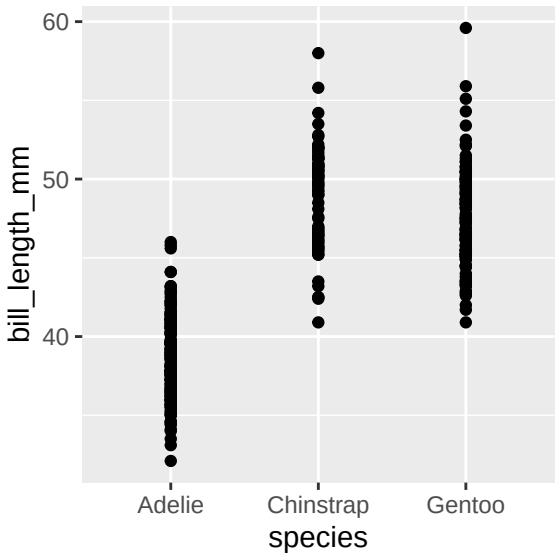
Using the same general structure, now plot the frequency distribution of penguin bill length, a *continuous variable*. Decide which of the following types of graphs would be the best way to display this distribution regardless of species:

- scatterplot: **geom_point()**
- violin plot: **geom_violin()**
- histogram: **geom_histogram()**

I think the box plot is best because it visually shows variance via quartiles, and it displays mean

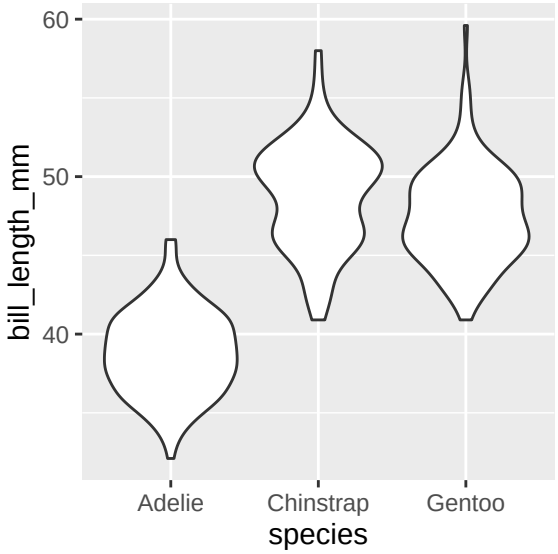
```
penguins %>% #piping in penguin data
#adding in bill length as response var, and species as explanatory var to plot
  ggplot(aes(y=bill_length_mm, x=species)) +
  geom_point() # making a scatter plot of the data
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```




```
penguins %>% #piping in penguin data
#adding in bill length as response var, and species as explanatory var to plot
  ggplot(aes(y=bill_length_mm, x=species)) +
  geom_violin() # making a violin plot of the data
```

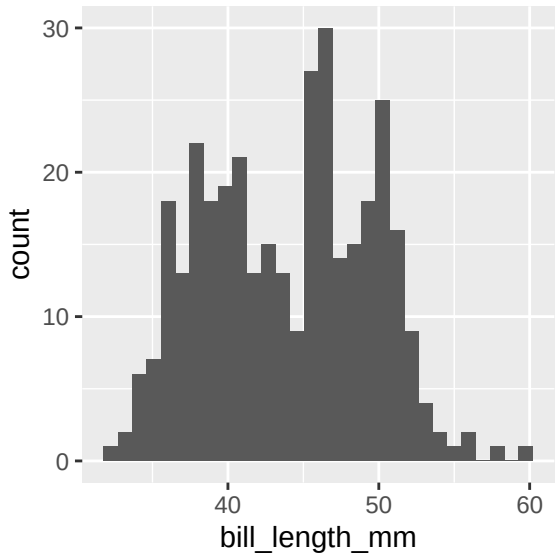
```
## Warning: Removed 2 rows containing non-finite values (stat_ydensity).
```




```
penguins %>% #pipint in penguin data
#adding in bill length as response var, and species as explanatory var to plot
  ggplot(aes(bill_length_mm)) +
  geom_histogram() # making a histogram of the data
```

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

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



You should have received a warning that 2 rows were removed because they contained "non-finite values". **What do you think this error means?** *Hint*: take a look at your output for **str(**penguins**)**.

I am guessing that the omitted rows have NA values.

Make a pretty plot!

We can make this graph a little bit nicer looking by doing the following:

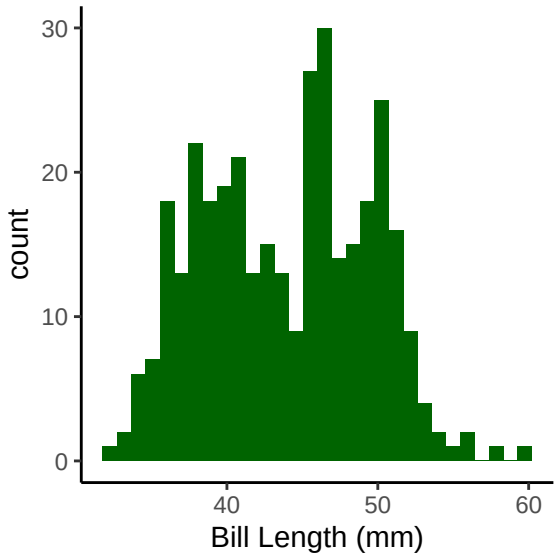
Remember, if you want to check what arguments a function needs, check the help file.

1. Renaming the x-axis using **xlab()** add-on
2. Changing the fill color of the histogram within **geom_histogram()**
3. Changing the overall theme of the graph using the **theme_classic()** add-on

```
penguins %>% #piping in the penguin data
  ggplot(aes(bill_length_mm)) + #adding bill length as the response var to plot
  geom_histogram(fill="darkgreen")+ #creating a green histogram of the data
  xlab("Bill Length (mm)")+ #labeling the axis
  theme_classic() #making the background of the fig white
```

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

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



As we discussed, this is a frequency distribution. Does this picture of the frequency of different bill lengths suggest any biological hypotheses?

It appears bill lengths of either around 40 or 50mm is optimal, but values in the middle are not advantageous. This is potentially an example of codominant traits.

Assignment

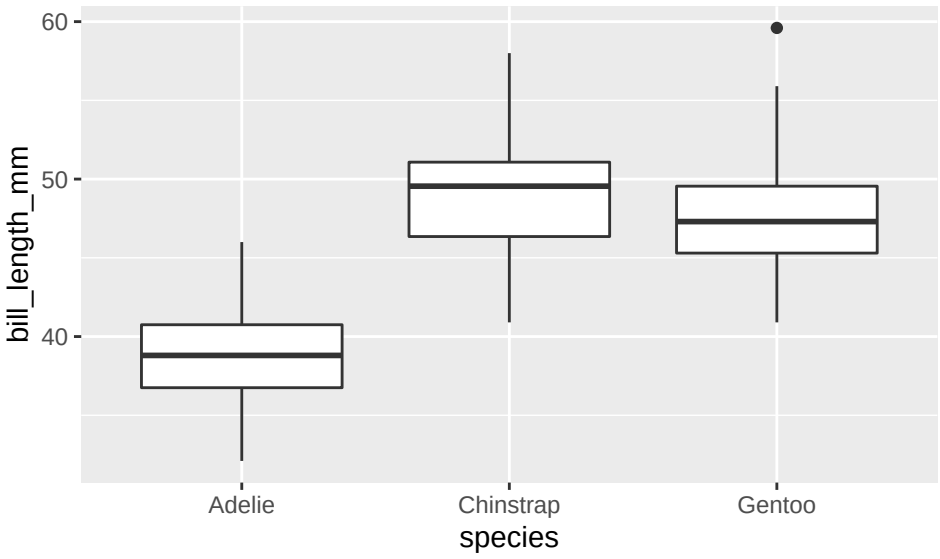
Question 1:

Does penguin bill length vary by species?

a. Create a boxplot (**geom_boxplot()**) that could be used to address that question. You will need both an x and a y variable in your aes() argument. Think about which variable should be the explanatory variable and which should be the response variable.

```
#making a box plot of penguin bill length by species
penguins %>% #piping in the dataset
  #assigning species as the explanatory variable and bill length as the response variable
  ggplot(aes(x=species, y=bill_length_mm)) +
  #making a basic box plot
  geom_boxplot()
```

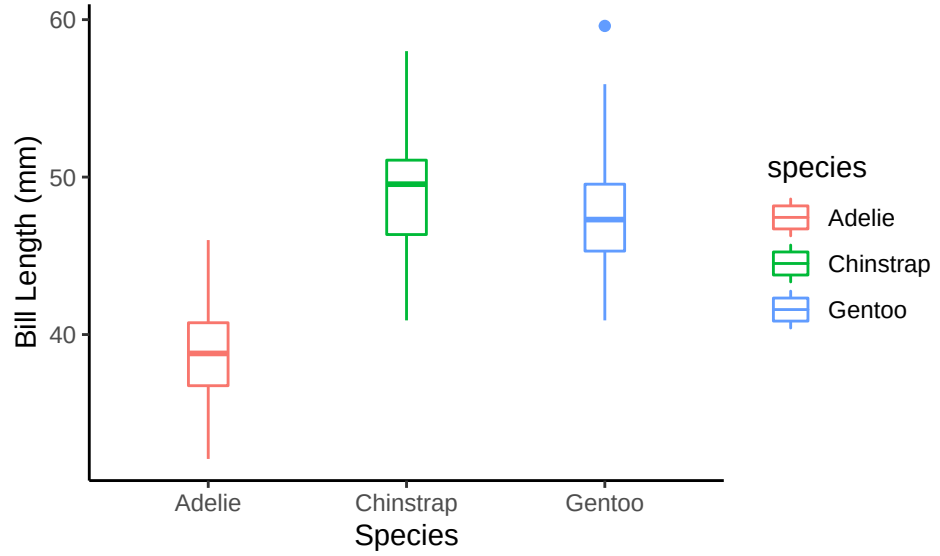
```
## Warning: Removed 2 rows containing non-finite values (stat_boxplot).
```



b. Make your plot pretty! This time, to adjust the color of each boxplot, we can assign color to our categorical variable of choice by putting "color = species" within the **aes()** argument.

```
penguins %>% #piping in the dataset
  #assigning species as the explanatory variable and bill length as the response variable, making each
  ggplot(aes(x=species, y=bill_length_mm, color=species)) +
  geom_boxplot(width=0.2) + #making boxplot of the data with width 0.2
  xlab("Species") + ylab("Bill Length (mm)") + #labeling x and y axis
  theme_classic() #setting theme to classic to change plot window background
```

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

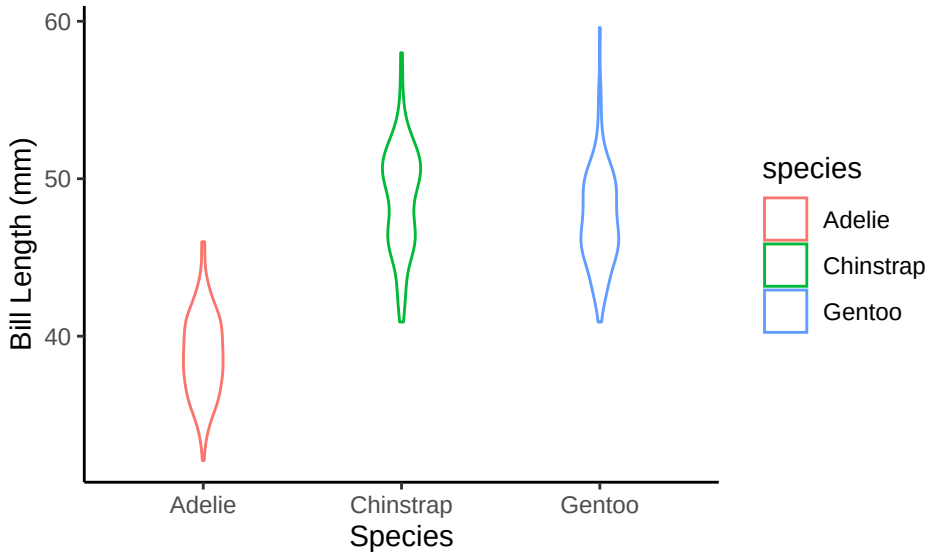


c. Create a violin plot (**geom_violin()**) that also could be used to address that question. **Write a sentence or two explaining which of the two plots you find more informative and why.**

The boxplot seems to be more informative about the difference in bill length among penguin species because it displays the mean and the 1st, 2nd, 3rd, and 4th quartiles, which to me makes comparing the spread of beak lengths more intuitive then trying to visually compare the frequency histograms in the violin plot.

```
penguins %>% #piping in the dataset
  #assigning species as the explanatory variable and bill length as the response variable, making each
  ggplot(aes(x=species, y=bill_length_mm, color=species)) +
  geom_violin(width=0.2) + #making boxplot of the data with width 0.2
  xlab("Species") + ylab("Bill Length (mm)") + #labeling x and y axis
  theme_classic() #setting theme to classic to change plot window background
```

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



d. **Do you think based on these plots that average penguin bill length varies by species? If so, which species appear to have different bill lengths?**

It appears that there is a significant difference in bill length among penguin species. In specific, the bill length of Adelie penguins appears to be significantly different from Chinstrap and Gentoo penguins (in this case shorter). Gentoo and Chinstrap penguins seem to have no significant difference in bill lengths.

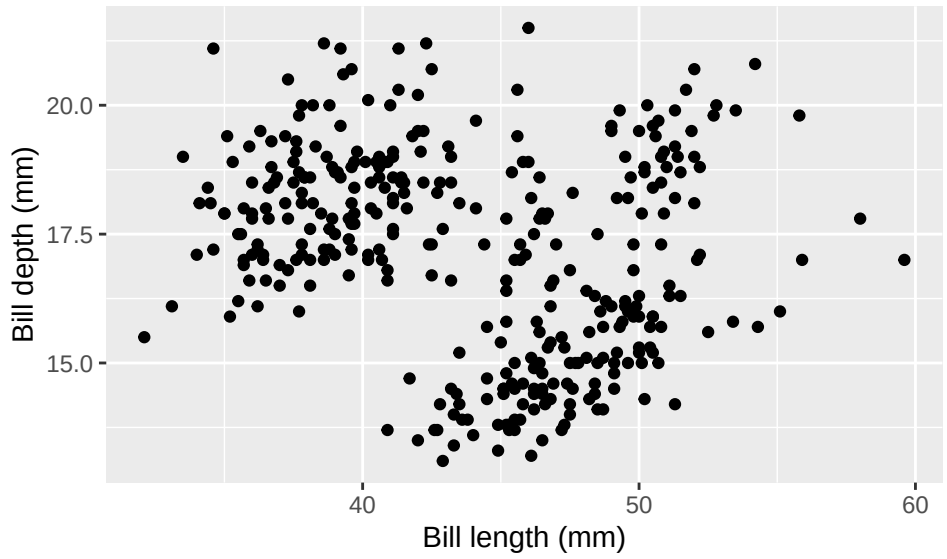
Question 2:

What is the relationship between bill length and bill depth?

a. Create a scatterplot (**geom_point()**) that describes the relationship between bill length and bill depth. Use bill depth as your response variable. Don't add color quite yet.

```
penguins %>% #piping the penguin data in
#adding bill length as explanatory var and bill depth as response variable to plot
  ggplot(aes(x=bill_length_mm,y=bill_depth_mm)) +
  geom_point() + #creating scatterplot of the data
  xlab("Bill length (mm)") + ylab("Bill depth (mm)") # adding axis labels
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



b. **Does there appear to be a relationship between these two variables? If so, is it positive or negative based on this graph alone?**

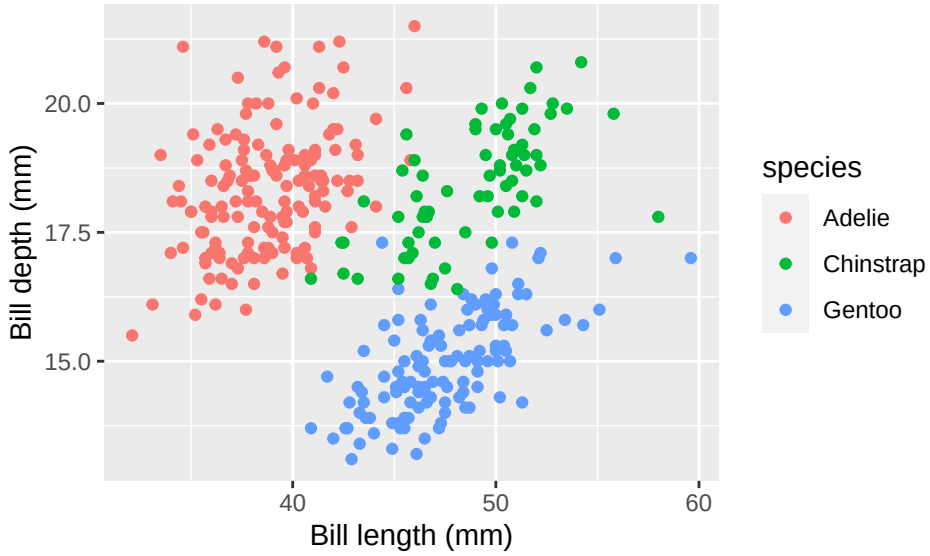
There appears to be either no relationship between these variables, or a very noise week positive correlation between the two variables.

c. Now color the points by species using the same methodology as in Question 1b. **How does this change your answer for 2b?**

Coloring points by species shows that within a given species, there seems to be a clear positive relation between bill length and depth in each given species. Color made this much easier to visualize.

```
penguins %>% #piping the penguin data in
#adding bill length as explanatory var and bill depth as response variable to plot, plus color as the sp
  ggplot(aes(x=bill_length_mm,y=bill_depth_mm,color=species)) +
  geom_point() + #creating scatterplot of the data
  xlab("Bill length (mm)") + ylab("Bill depth (mm)") # adding axis labels
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



Question 3:

Fill in the comments (# with text afterwards within the code chunk) to break down what is going on in this code, line by line, based on the plot that is produced from the code. Use Help within RStudio or type the function name with a question mark before hand in the console to learn more about the arguments (*e.g.* **?geom_point()**). You can also use the ggplot cheatsheet (linked **here**).

```
penguins %>% #piping in penguins data
#creating plot with the response variable as bill length, explanatory var as island the penguin was liv
  ggplot(aes(x = island, y = bill_length_mm)) +
#making a "jittered" scatter plot of the data with the amount of vertical and horizontal jitter/stagger
  geom_jitter(height = 0, width = 0.2,
              size = 2, alpha = 0.4) +
#adding purple "crossbar" line graphic  element 0.5 units wide at the mean of each island's data
  stat_summary(fun = mean, geom = "crossbar",
                 width = 0.5, color = "purple") +
  ylab("bill length (mm)") +
  theme_minimal() #chaning background color of plot to "minimal" preset
```

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

```
## Warning: Removed 2 rows containing missing values (geom_point).
```