SOLUTIONS LAB ACTIVITY: R Workshop Day 1

Categorial Variables, Scatterplots, Correlation

Eric Carreira

2023-05-18

There is a data frame called **penguins** inside a package called **palmerpenguins.** In the console, you can use the view() function to open this data frame in a new tab.

Please insert code chunks below each question.

Use the **names()** function to list all the variables in this data.

```
names(penguins)
```

```
## [1] "species"           "island"            "bill_length_mm"
## [4] "bill_depth_mm"     "flipper_length_mm" "body_mass_g"
## [7] "sex"               "year"
```

PART 1. CATEGORICAL VARIABLES

In your view(penguins) tab, see that there are two categorical variables: species and island.

SPECIES
Use the **tally()** function to count how many penguins of each species are in this data.

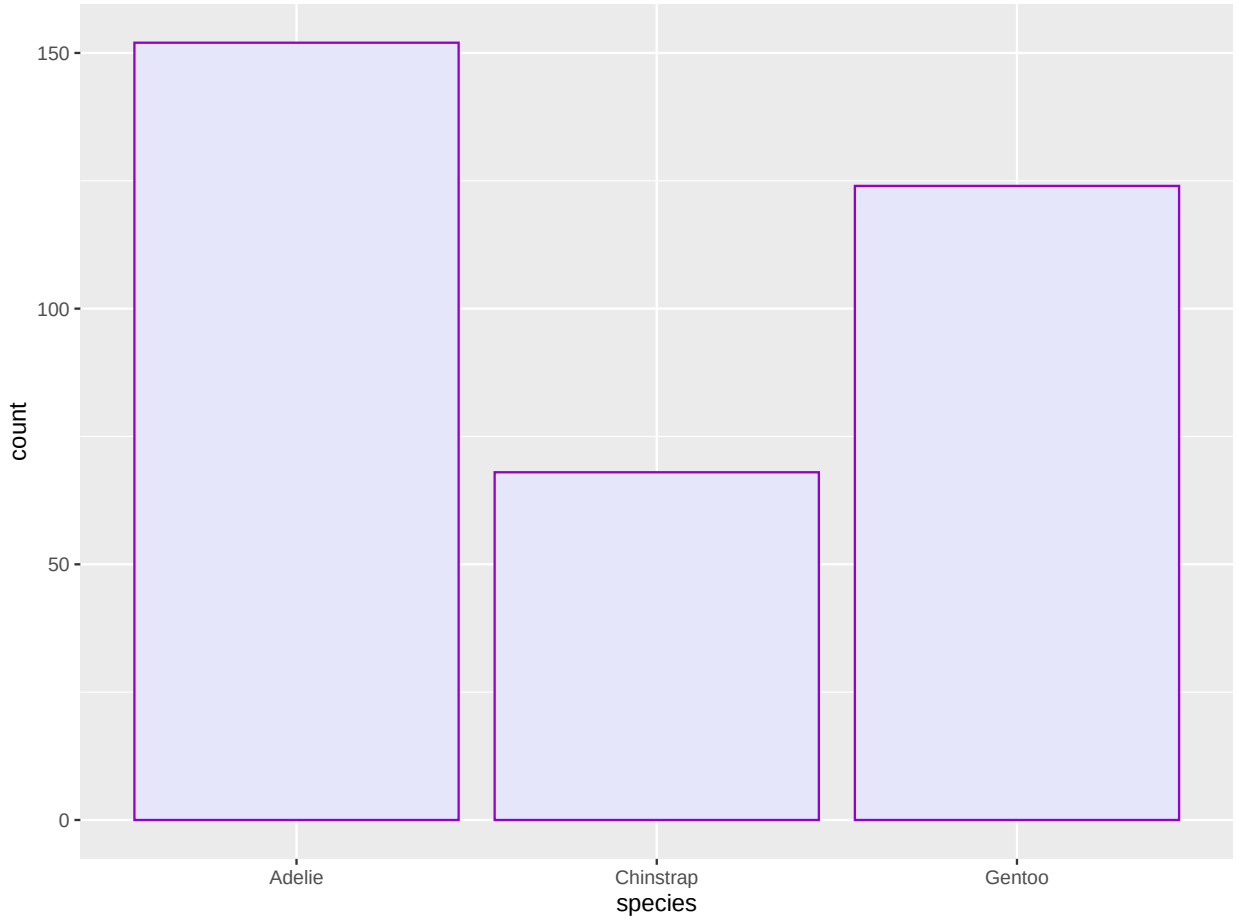
```
tally(~ species, data = penguins)
```

```
## species
##    Adelie Chinstrap    Gentoo
##       152        68       124
```

Let's make visualizations of the above table.

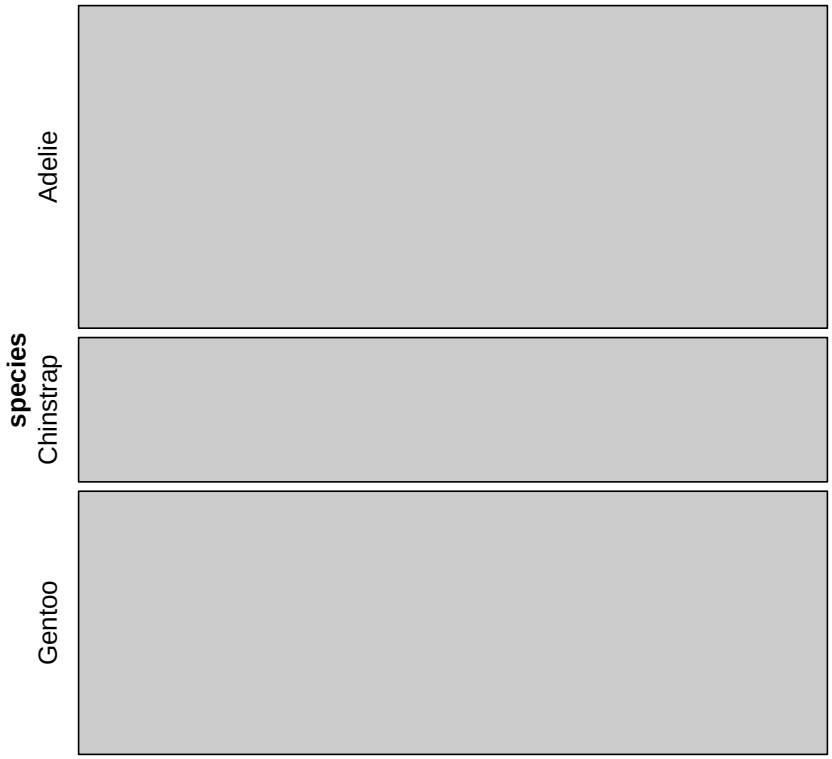
Use **gf_counts()** to make a bar plot of the different species. Use color and fill commands to make it more visually appealing!

```
gf_counts(~ species, data = penguins, color = "darkviolet", fill = "lavender")
```



Use the **mosaic()** function to make a mosaic plot of the species variable.

```
mosaic(~ species, data = penguins)
```



ISLAND
Use the **tally()** function to find out the penguin count on each island for this data set. *BE CAREFUL! Do not click "islands" suggestion! That is the name of different data set that RStudio is recognizing!*

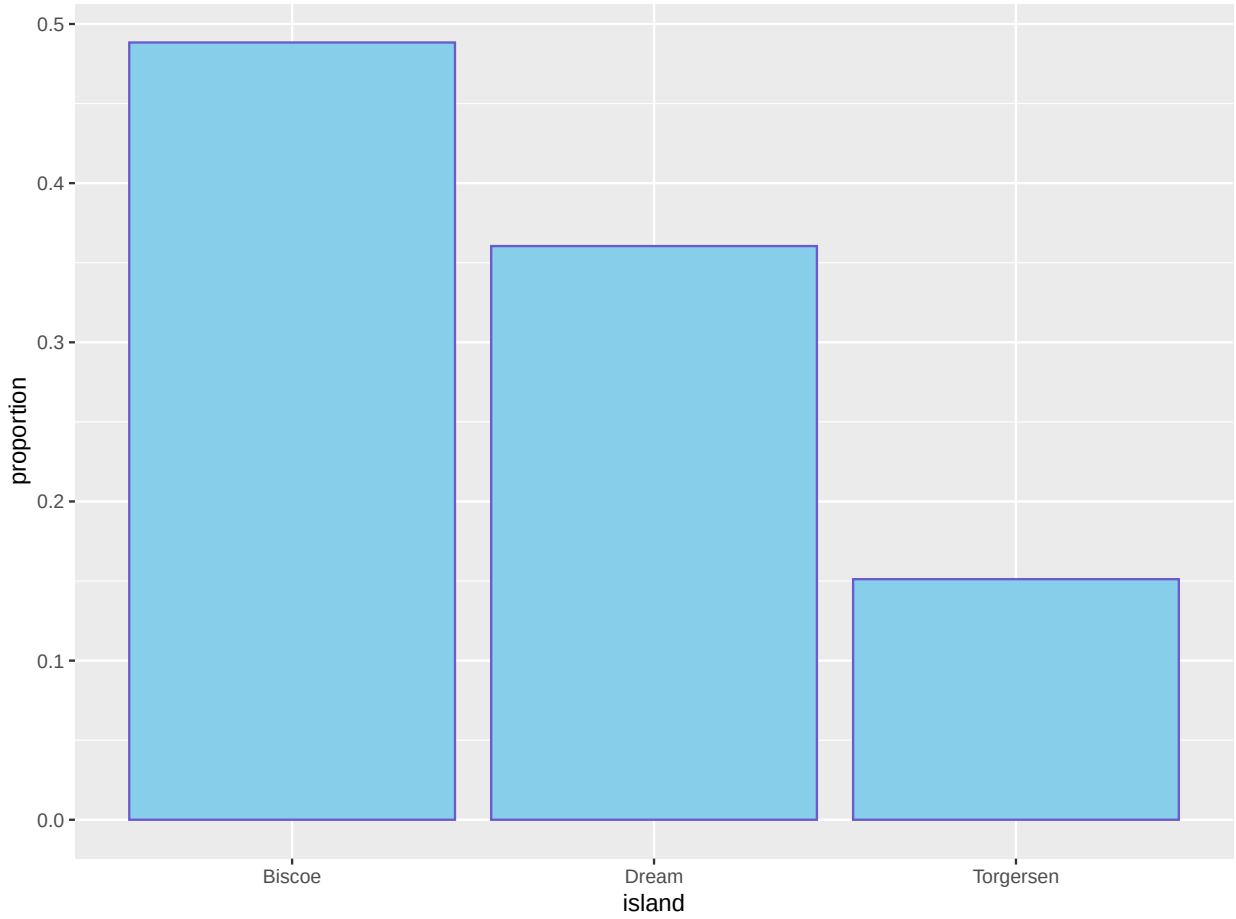
```
tally(~ island, data = penguins)
```

```
## island
##    Biscoe     Dream Torgersen
##       168       124        52
```

Let's make visualizations of the above table.

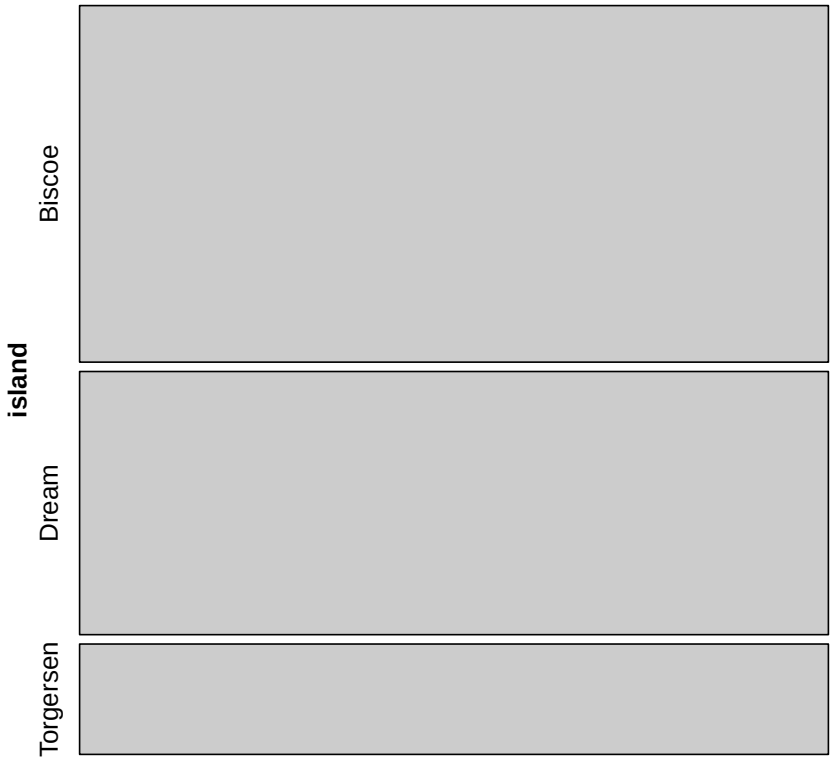
Use **gf_props()** to make a bar plot of the island variable. Use color and fill commands to make it more visually appealing!

```
gf_props(~ island, data = penguins, color = "slateblue", fill = "skyblue")
```



Use the **mosaic()** function to make a mosaic plot of the island variable.

```
mosaic(~ island, data = penguins)
```



SPECIES ON EACH ISLAND

Use the tally() function to count the number of each different species on each of the different islands. (This is a *2-by-2 contingency table!*)

```
tally(~ species + island, data = penguins)
```

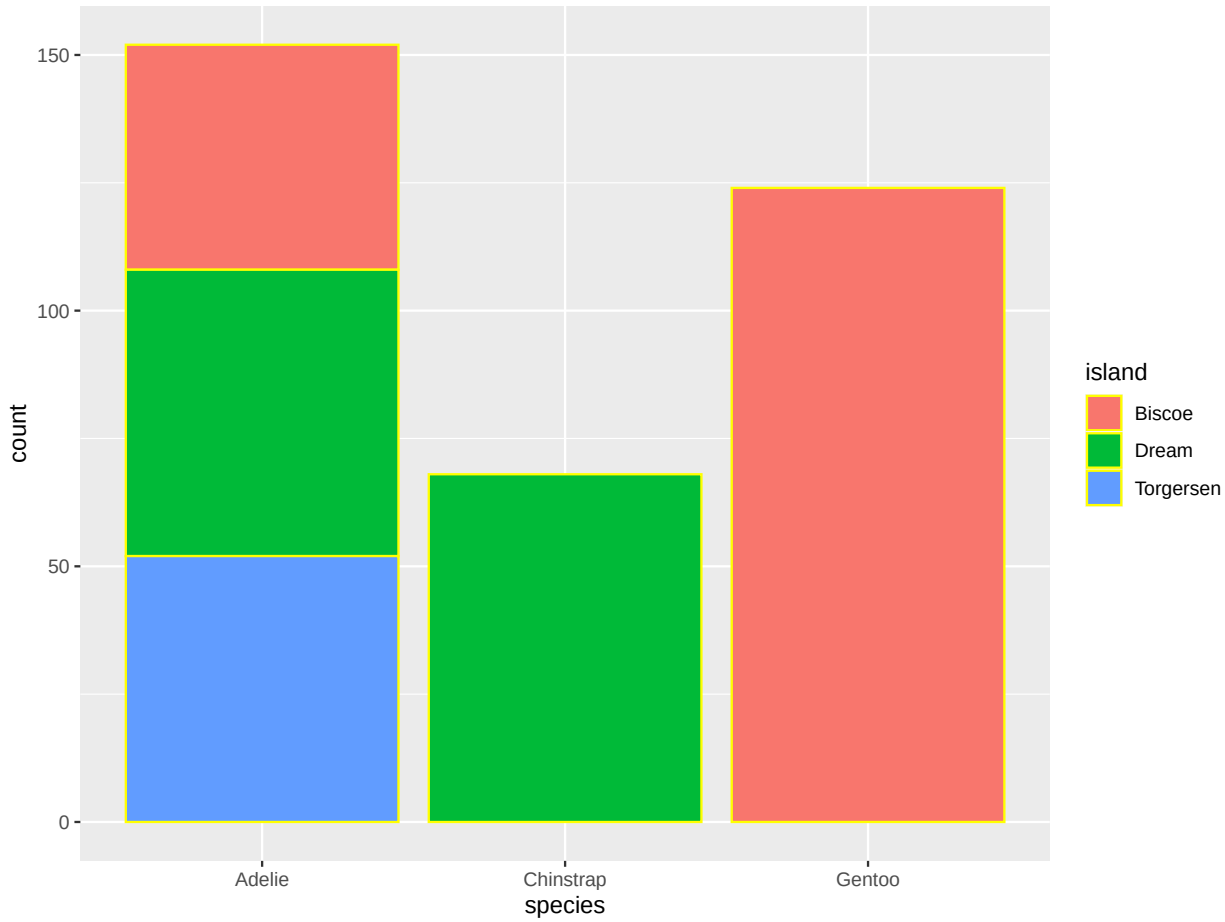
```
##            island
## species     Biscoe Dream Torgersen
##   Adelie        44    56        52
##   Chinstrap      0    68         0
##   Gentoo       124     0         0
```

Intepret the values in this table. Which species/island combination is most represented in this data? In words, what do the zeros mean?

Let's make visualizations of the above table.

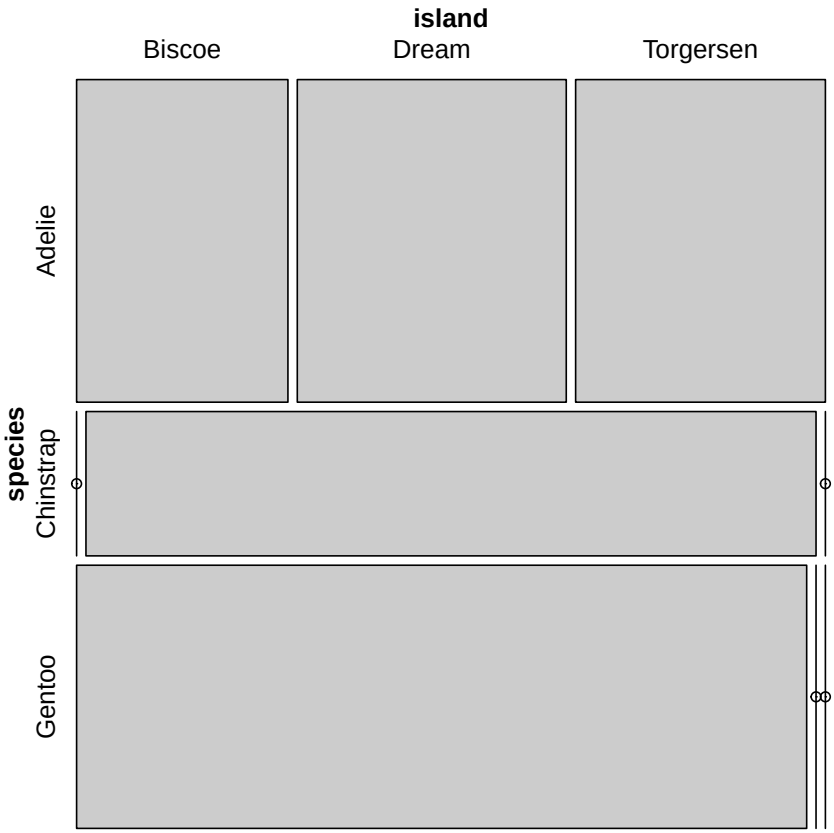
Use **gf_percents()** to make a bar plot of the different species and the fill = to color each bar according the island variable. This allow us to see for each species, how many came from each island.

```
gf_counts(~ species, data = penguins, color = "yellow", fill = ~ island)
```



Use the **mosaic()** function to make one mosaic plot of the species and island variables together.

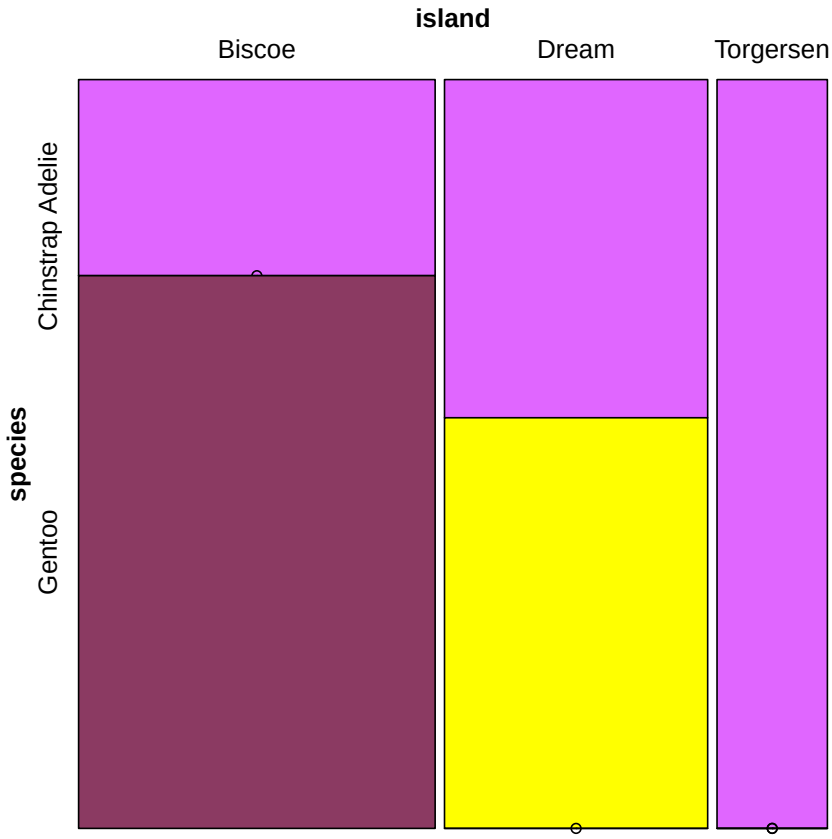
```
mosaic(~ species + island, data = penguins)
```



For an extra challenge, let's try adding color.
Use the highlighting argument to specify a variable to highlight with variable name in quotes.
For example, highlighting = "species".
Use the highlighting_fill argument to specify your colors. In this case, three colors are needed.
For example, highlighting_fill = c("color1", "color2", "color3") Insert your color names inside quotes!

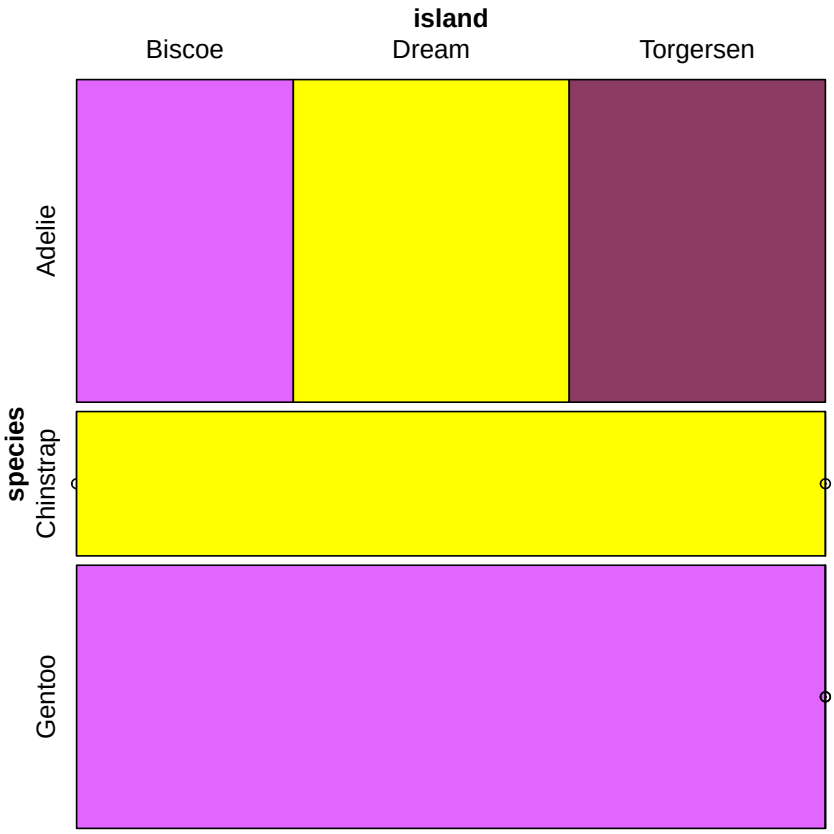
Highlight By Species:

```
mosaic(~ species + island, data = penguins,
       highlighting = "species",
       highlighting_fill = c("mediumorchid1", "yellow", "hotpink4"))
```



Highlight By Island:

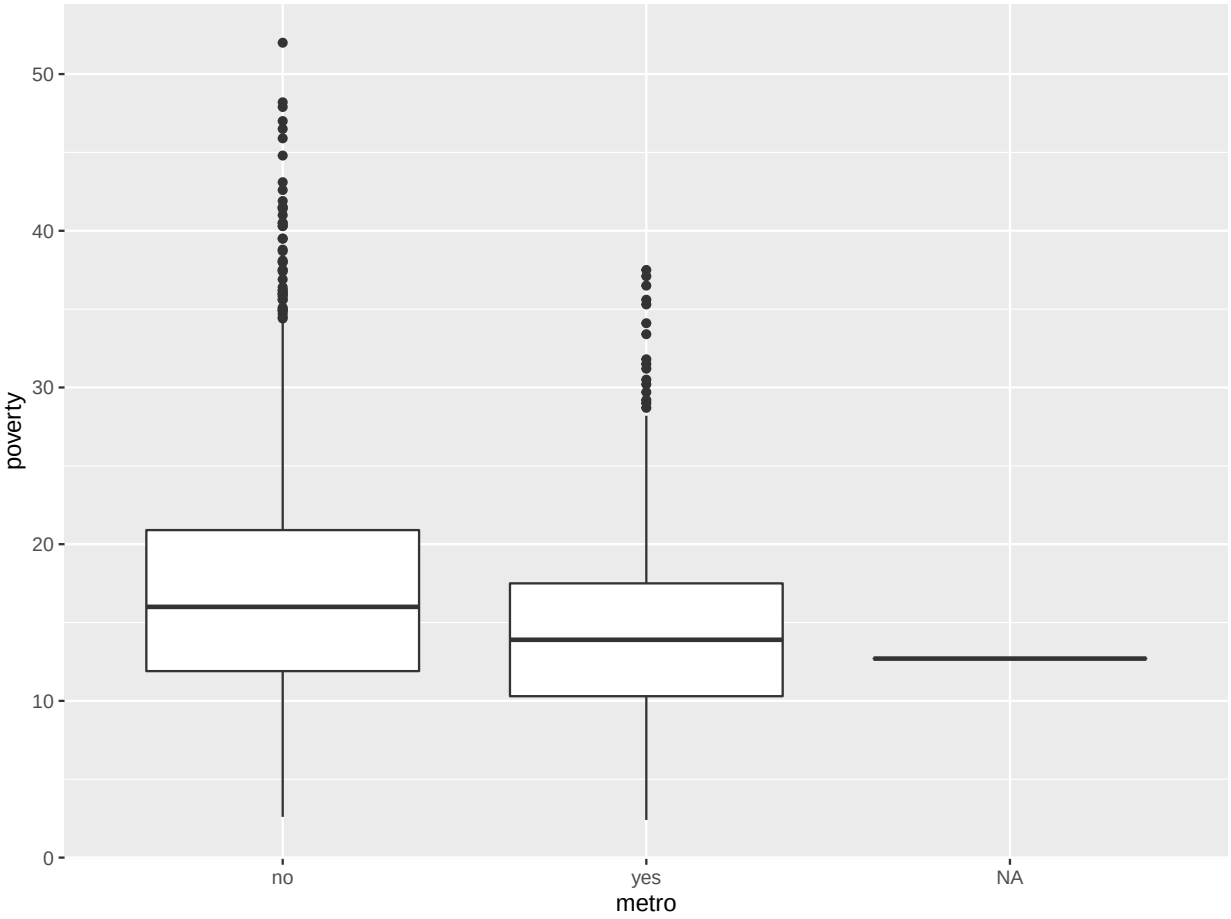
```
mosaic(~ species + island, data = penguins,
       highlighting = "island",
       highlighting_fill = c("mediumorchid1", "yellow", "hotpink4"))
```



PART 2. ONE NUMERICAL, ONE CATEGORICAL VARIABLE

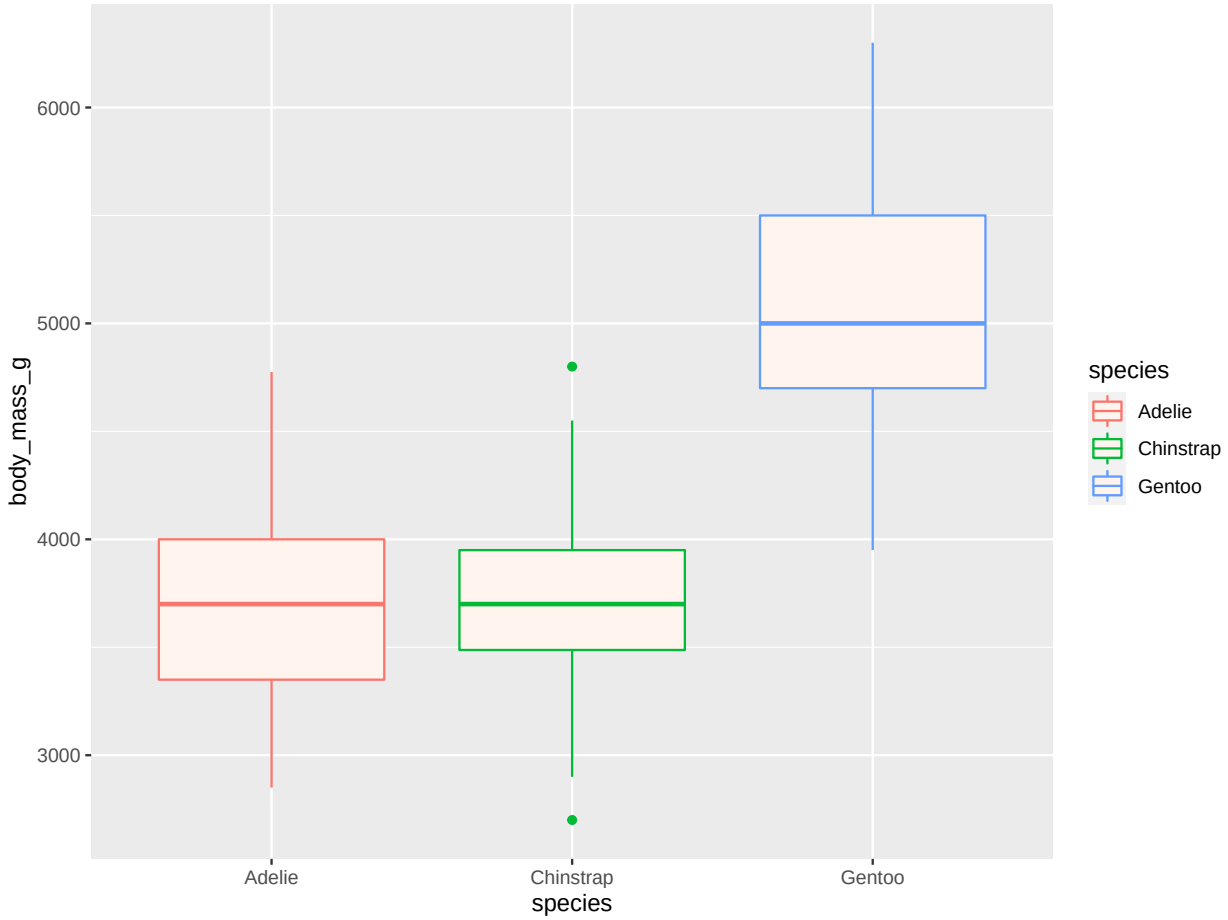
Recall from class how we made comparative side-by-side boxplots using the **county** data. We did this by taking the numerical variable of poverty rates (called **poverty**) separating the poverty rates into two groups based on the "yes or no" levels of the categorical variable **metro**. Recall that metro was "yes" if county has a large urban metro area) or "no" (county does not have a large urban metro area.)
The code we used for this is given below:

```
gf_boxplot(poverty ~ metro, data = county)
```



Use this same idea to make comparative side-by-side boxplots of the numerical variable called **body_mass_g** which gives the weight of the penguins in grams.

```
gf_boxplot(body_mass_g ~ species, data = penguins, color = ~ species, fill = "seashell")
```



Which penguin species has the largest median weight? Did any species have outliers? Which species has the smallest IQR (DO you remember how to see IQR in a boxplot?) Largest median weight is Gentoo. Only Chinstrap has outliers (one low, one high). Chinstrap has the smallest IQR as indicated by the narrowed box.

Use the same technique above with **favstats()** to find all the major summary statistics of **body_mass_g** for all three species using just one line of code! Be careful of the order of the variables! What happens if you reverse them? (Notice in the favstats output that there is missing data!)

```
favstats(body_mass_g ~ species, data = penguins)
```

```
##     species  min     Q1 median   Q3  max     mean       sd   n missing
## 1    Adelie 2850 3350.0   3700 4000 4775 3700.662 458.5661 151       1
## 2 Chinstrap 2700 3487.5   3700 3950 4800 3733.088 384.3351  68       0
## 3    Gentoo 3950 4700.0   5000 5500 6300 5076.016 504.1162 123       1
```

Use the same idea above to compute IQR for each species. You need to use an argument to remove missing data.

```
iqr(body_mass_g ~ species, data = penguins, na.rm = TRUE)
```

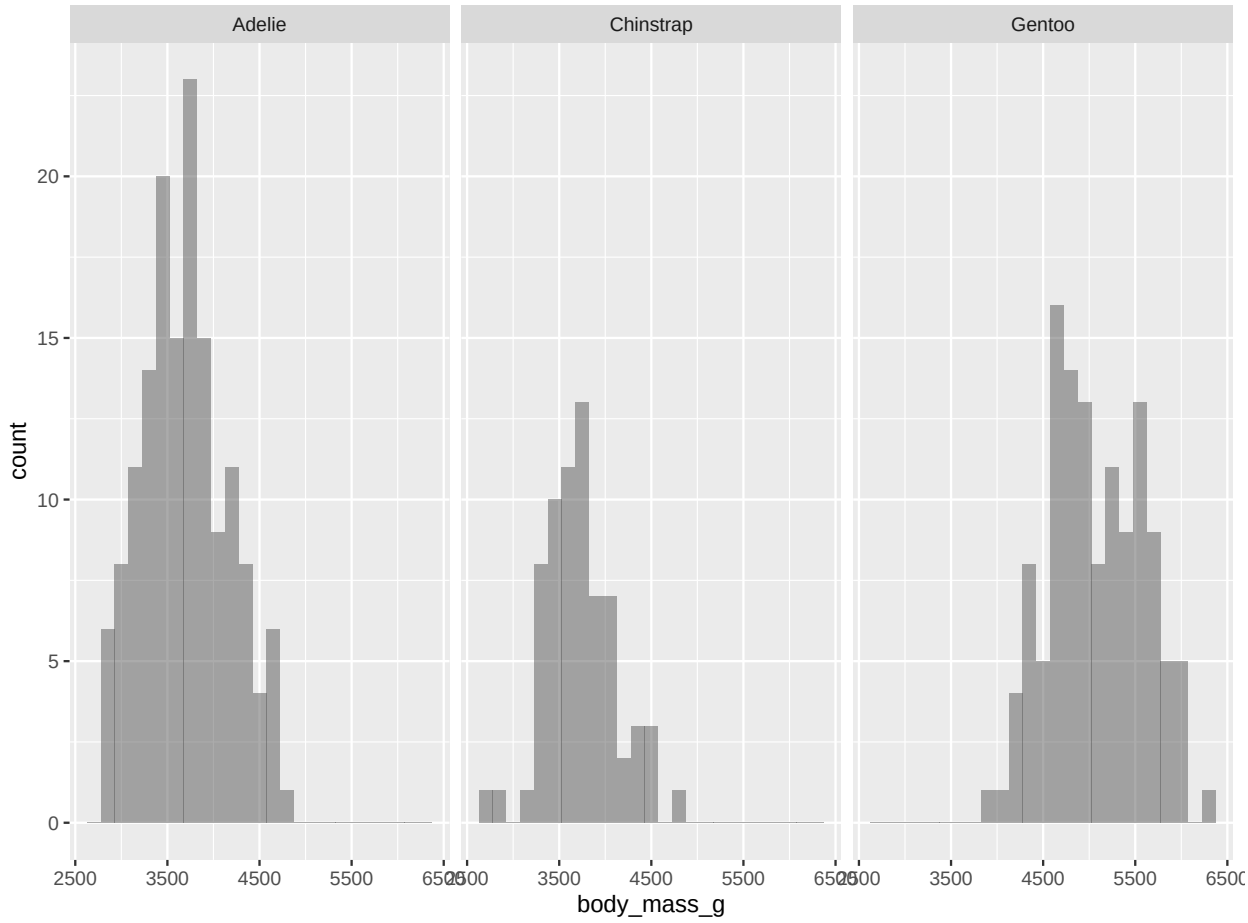
```
##    Adelie Chinstrap    Gentoo
##     650.0     462.5     800.0
```

Which species has the least variability in the middle 50%? Does this agree with what you saw in the side-by-side boxplots?
Chinstrap has the smallest IQR just as the boxplots indicated.

Make a histogram of body mass faceted by species and also colored by species.
For faceting, remember to use the "|" meaning body_mass_g broken down by species.
Alternatively, pipe %>% to the mosaic function gf_facet_wrap() and put species inside.

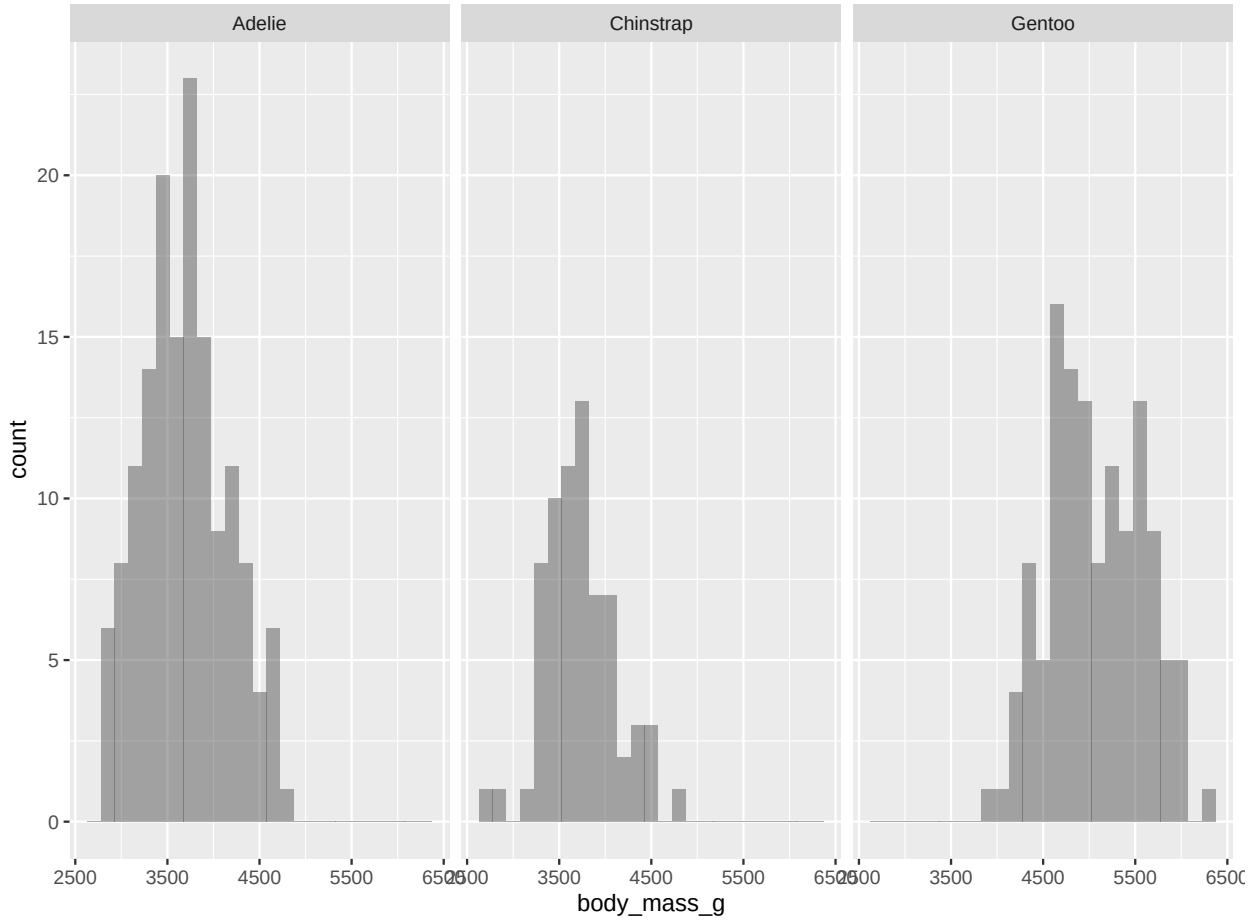
Option 1

```
gf_histogram(~ body_mass_g | species, data = penguins)
```



Option 2

```
gf_histogram(~ body_mass_g, data = penguins) %>% gf_facet_wrap(~ species)
```



PART 3. TWO NUMERICAL VARIABLES

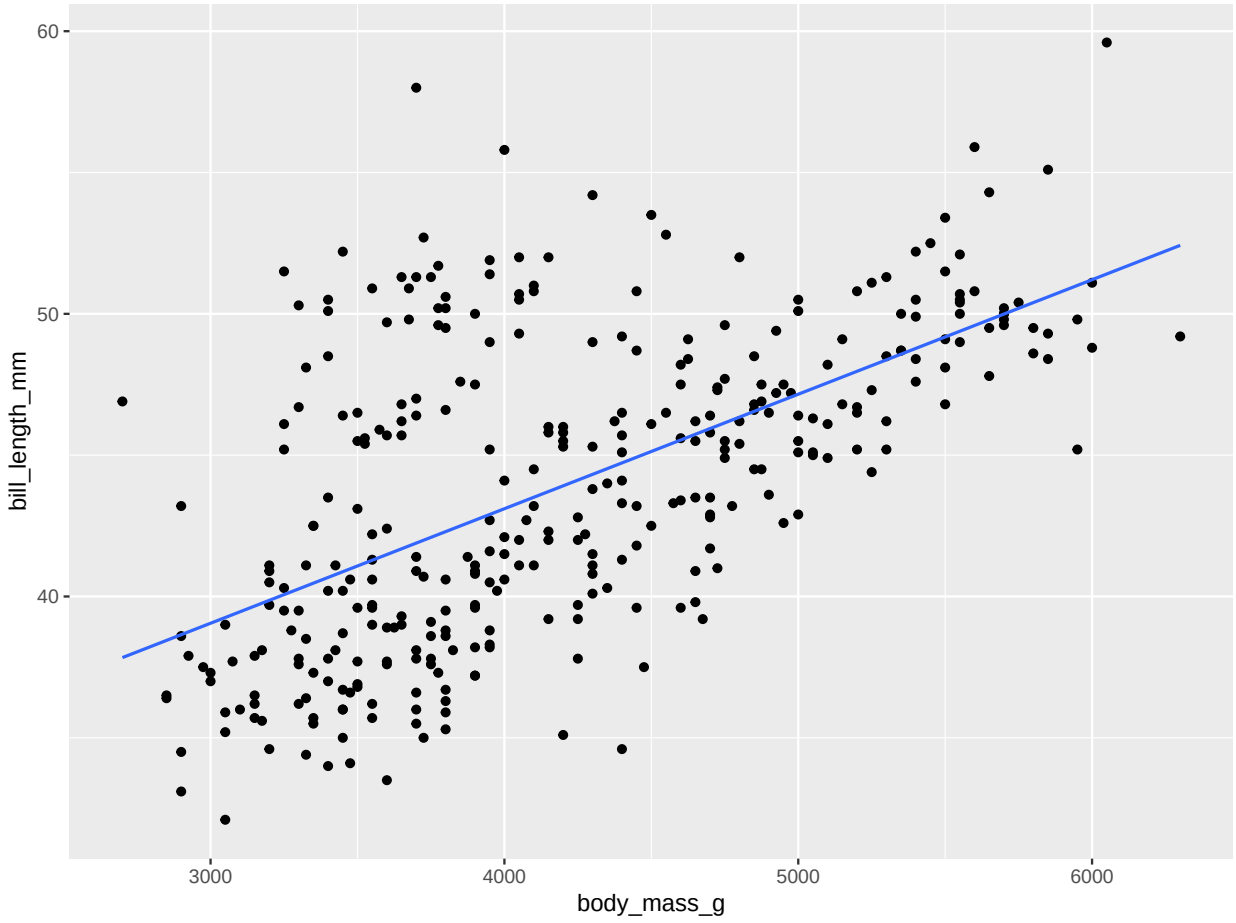
SCATTERPLOTS AND CORRELATION

It seems plausible that, regardless of species, penguins with larger body mass could also have longer flippers and a longer bill. Let's make scatterplots for each of these variables to examine their relationship with body mass. And from the graphs, we will consider which of these two variables seems to be more highly correlated with body mass.

First, make a scatterplot with bill length on the y-axis, body mass on the x-axis. (Make sure you get the correct variable names! Use the names() function or go look at the data frame! And remember y-axis variable goes first inside the function as y ~ x). Be sure to include the straight trend line to your plot!

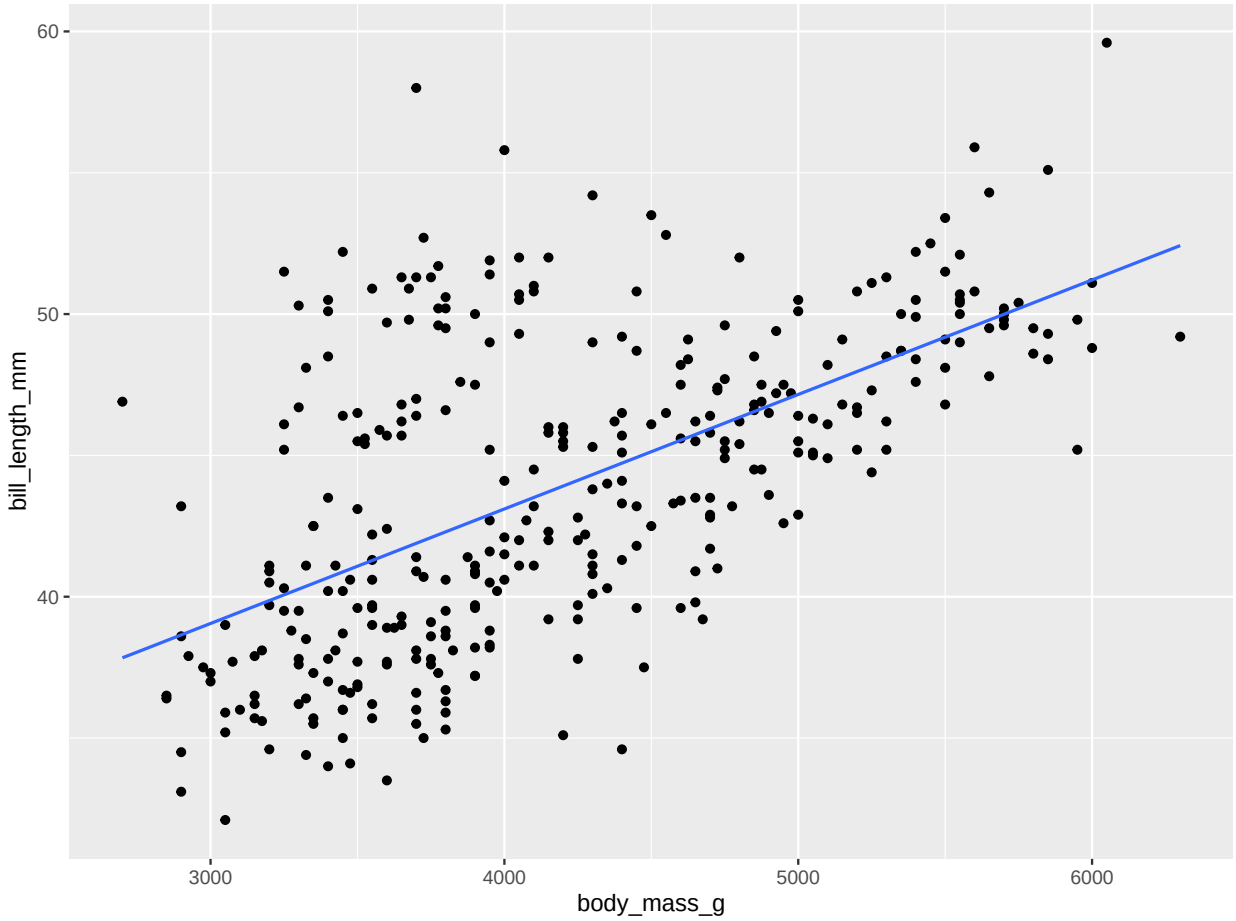
Option 1

```
gf_point(bill_length_mm ~ body_mass_g, data = penguins) + geom_lm()
```



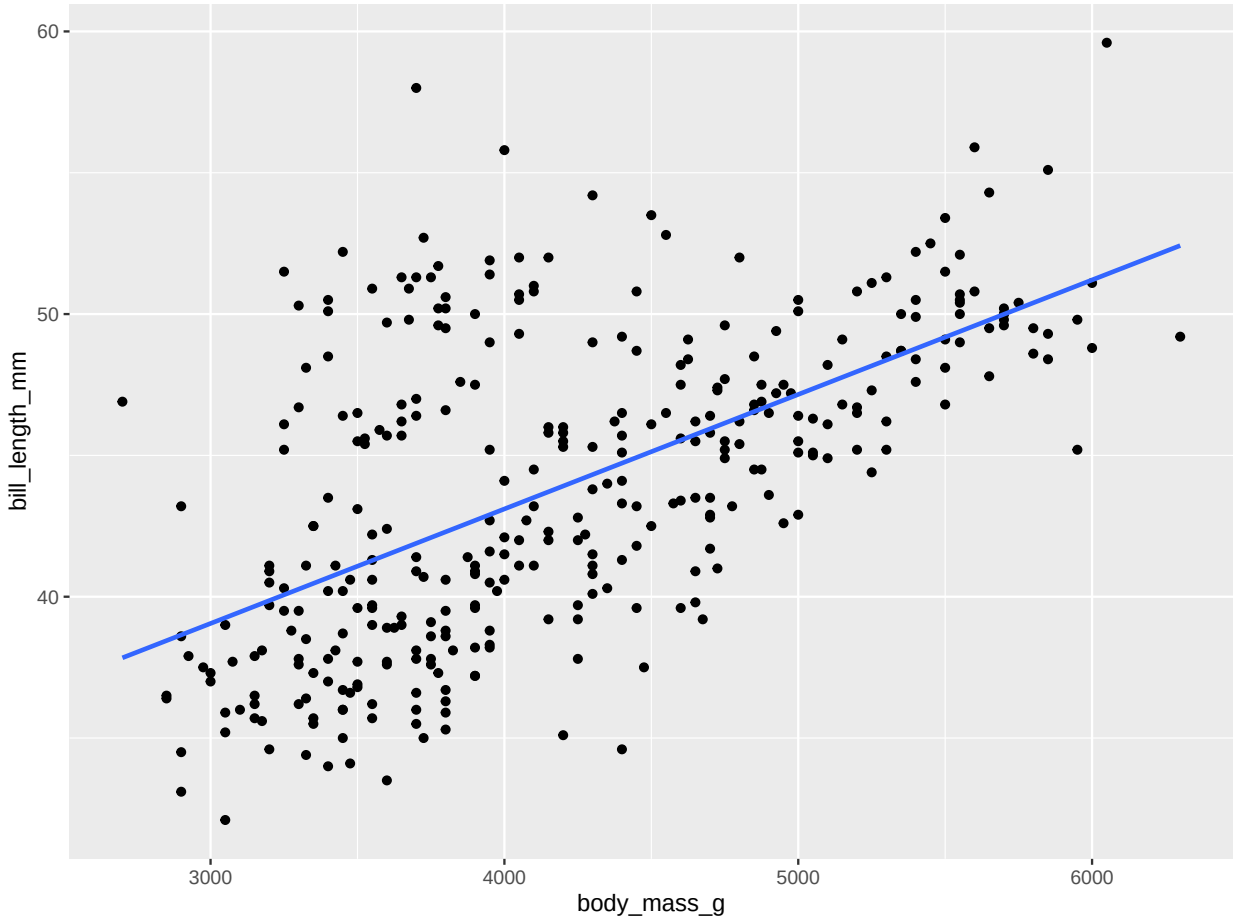
Option 2

```
gf_point(bill_length_mm ~ body_mass_g, data = penguins) %>% gf_lm()
```



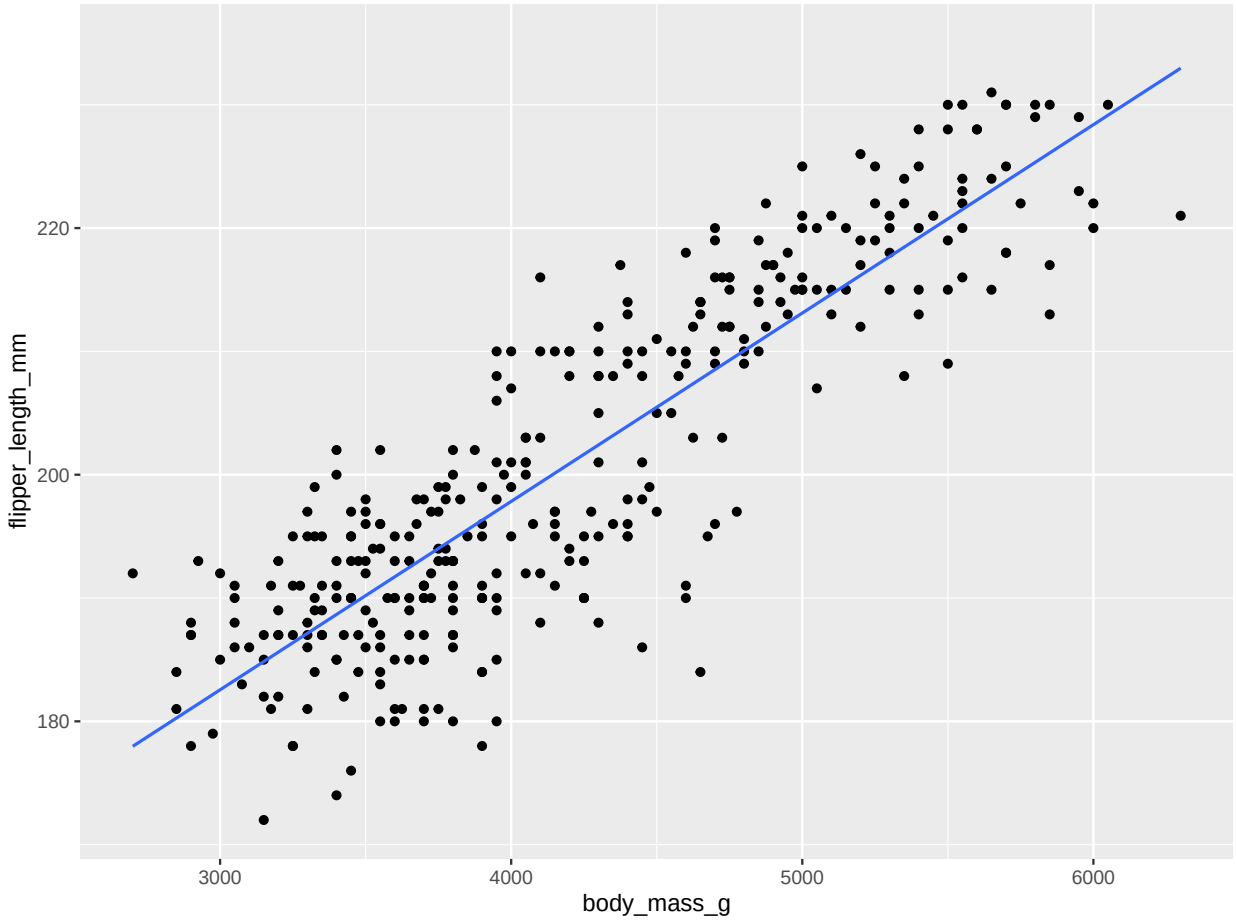
Option 3

```
gf_point(bill_length_mm ~ body_mass_g, data = penguins) %>% gf_smooth(method = "lm")
```



Next, make a scatterplot with flipper length on the y-axis, body mass on the x-axis. Be sure to include the straight trend line to your plot with *__lm().

```
gf_point(flipper_length_mm ~ body_mass_g, data = penguins) + geom_lm()
```



From these two scatterplots, do you see a linear relationship? You should! Does bill length or flipper length appear to be more highly correlated with body mass or do they look about the same? In other words, do the points appear more closely clustered around the straight line in one of the scatterplots?

Let's dig deeper into this question by finding the value of the correlation statistic. There is missing data so include the command **use = "complete"**

Compute the Correlation statistic for bill length and body mass.

```
cor(bill_length_mm ~ body_mass_g, data = penguins, use = "complete")
```

```
## [1] 0.5951098
```

Compute the Correlation statistic for flipper length and body mass.

```
cor(flipper_length_mm ~ body_mass_g, data = penguins, use = "complete")
```

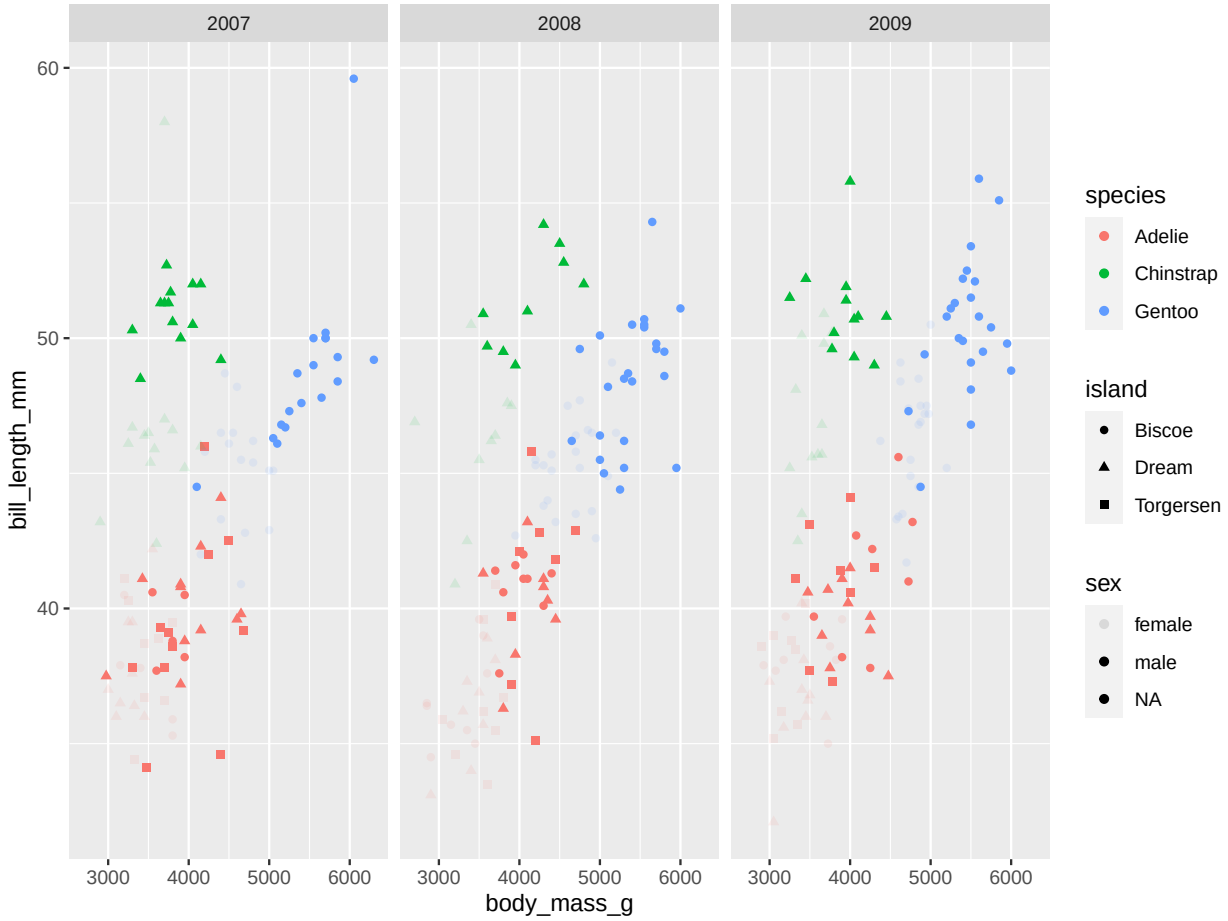
```
## [1] 0.8712018
```

PART 4. FUN CHALLENGE!

MULTIVARIABLE SCATTERPLOT

Choose one of the scatterplots above and try to map as many variables as you can to the plot. Can you get 6 of the 8 variables onto the plot?

```
gf_point(bill_length_mm ~ body_mass_g | year, data = penguins,
         color = ~ species,
         alpha = ~ sex,
         shape = ~ island)
```



PART 5. OPTIONAL HOMEWORK!

1) Finish the activity above if you have not done so already.

AND/OR

2) Repeat some of the same analyses on a different data frame as decribed below.

In the **mdsr** package, there is a data frame called HELPrct. The HELP study was a clinical trial for adult inpatients recruited from a detoxification unit. In the console, type ?HELPrct to learn more.

Use categorical variables substance, homeless, sex and numerical variables cesd (measure of depressive symptoms, higher scores indicate more symptoms) and mcs (mental component score, higher scores indicate better functioning). Experiment with tally, bar plots, mosaic plots, scatterplots of cesd and mcs, correlation, etc.