Final Project

Ellen Bledsoe

2022-12-03

Final Project Details

Purpose

The goal of this final assignment is to assess your ability to integrate the many skills you have learned over the semester: filtering and summarizing data, creating new columns, choosing the appropriate data visualizations, and performing and interpreting the appropriate statistical tests.

Task

Write R code which produces the correct data, summaries, plots and analyses. Correctly interpret the results of these plots and analyses.

Criteria for Success

- Code chunks run without errors
- Code produces the correct result
  - Code that produces the correct answer will receive **full** credit
  - Code attempts with logical direction will receive **partial** credit
- Appropriate plot types are used to visualize the data
- Appropriate statistical tests are used to analyze the data
- Written answers address the questions in sufficient detail

Due Date

December 12 at midnight MST

Final Project

For your final "project" this semester, I am presenting you with 3 problem sets, each worth 20 points (for a total of 60 points).

I'm expecting you to be able to filter and summarize the data in ways you need, choose the appropriate visualization, choose the appropriate analysis, and correctly interpret the analysis for the question I've asked you.

It is important to note that I will not be giving you an answer key for this final project since part of your grade for the assignment is being able to choose the appropriate visualizations and analyses for the question and the data.

We are going to use the `palmerpenguins` R package and data set, which we've used many times before! You can learn more about it here. This is a real data set from a Long-Term Ecological Research (LTER) site in Antarctica.

Set-Up

Be sure to run both of these code chunks before you begin! I've gone ahead and included the code to load the two packages you will need to successfully complete this project: the `tidyverse` and `palmerpenguins`. Be sure to run this code chunk!

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(palmerpenguins)
```

Important! I've also included one more code chunk below. Be sure to run this code chunk, as well! It does two key things:

- first, it add the `penguins` data frame to your environment, which I imagine you will find helpful
- second, it removes all rows that have any NA values, which will make completing this assignment a bit easier.

Once you've run this line of code, you should see the `penguins` data frame pop up in your environment with 333 observations (rows) and 8 variables (columns).

```
penguins <- penguins %>% drop_na()
```

Structure & Guidelines

Like the practice version, this final project is structured as 3 different problem sets. For each problem set, I am presenting you with a initial question to guide your thinking and analysis.

Data

Assume that nothing carries over between problem sets.

Each problem set is stand-alone, meaning that you should always start with the `penguins` data frame at the beginning of each problem set. If you should use a data frame that you created *within* the problem set, I explicitly state so.

For example, in Problem Set 2, you should use the `biscoe` data frame that you create for the entire problem set; at the start of Problem Set 3, start over with the `penguins` data frame.

Interpreting Statistical Results

When I ask you to interpret statistical results, you should roughly follow these guidelines.

- the cut-off for our p-values is always 0.05
- report the p-value that we are focused on
- if there are multiple p-values of interest, report all of them
- state whether the p-value indicates a significant difference/relationship
- if applicable, state whether we should or should not reject the null hypothesis

Plotting

All plots should be made using `ggplot2`.

You options for plot types to choose from are:

- histogram (use transparency (`alpha`) and `position = "identity"` with multiple groups to see the full distributions)
- density (use transparency (`alpha`) with multiple groups to see the full distributions)
- box-and-whisker (adding points on top is encouraged but not required)
- scatter plot

Note: All plots should have modified axis labels and legend labels.

In many cases, this might mean capitalizing the axis label or legend label. In other cases, you might want to put units in parentheses after the words (e.g., Body Mass (g)).

Problem Set 1 (20 points)

Question: Are there differences in the average bill length across the 3 islands in the data set: Dream, Biscoe, and Torgersen? (Ignore species) Let's start by summarizing the bill length data.

1. Calculate the minimum, maximum, and mean bill length for *each* island. (3 points)

```
penguins %>%
  group_by(island) %>%
  summarise(min_bill = min(bill_length_mm),
            max_bill = max(bill_length_mm),
            mean_bill = mean(bill_length_mm))
```

```
## # A tibble: 3 x 4
##   island    min_bill max_bill mean_bill
##   <fct>        <dbl>    <dbl>     <dbl>
## 1 Biscoe        34.5     59.6      45.2
## 2 Dream         32.1     58        44.2
## 3 Torgersen     33.5     46        39.0
```

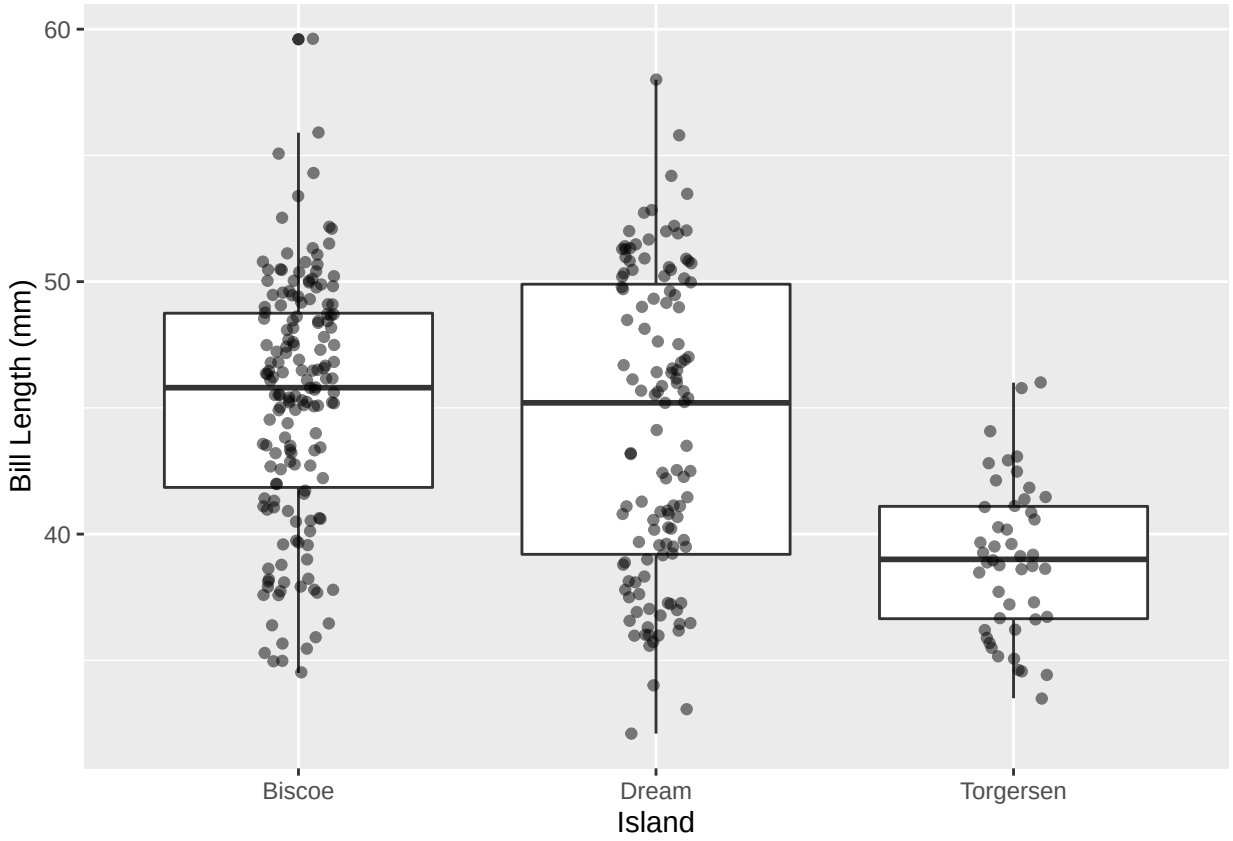
2. Which of our variables would be considered *independent* and which one *dependent*? Also determine whether each is *continuous* or *categorical*. (4 points)

    - **island**: independent, categorical
    - **bill length**: dependent, continuous

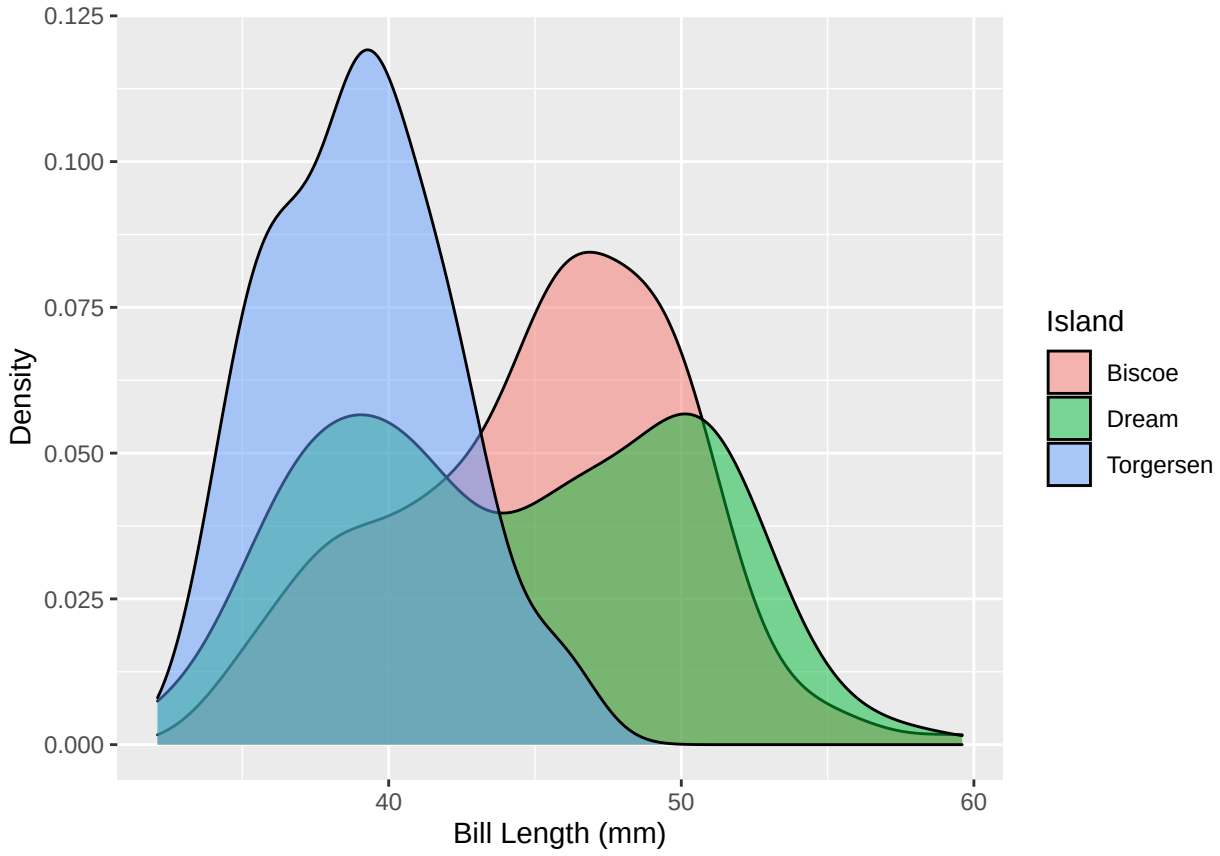
Now that we have an idea numerically of the differences between the islands, let's plot the differences.

3. Choose an appropriate plot (there are a few options). Ensure that you follow the plotting guidelines in the Structure & Guidelines section above! (3 points)

```
ggplot(penguins, aes(island, bill_length_mm)) +
  geom_boxplot() +
  geom_jitter(width = 0.1, alpha = 0.5) +
  labs(x = "Island",
       y = "Bill Length (mm)")
```

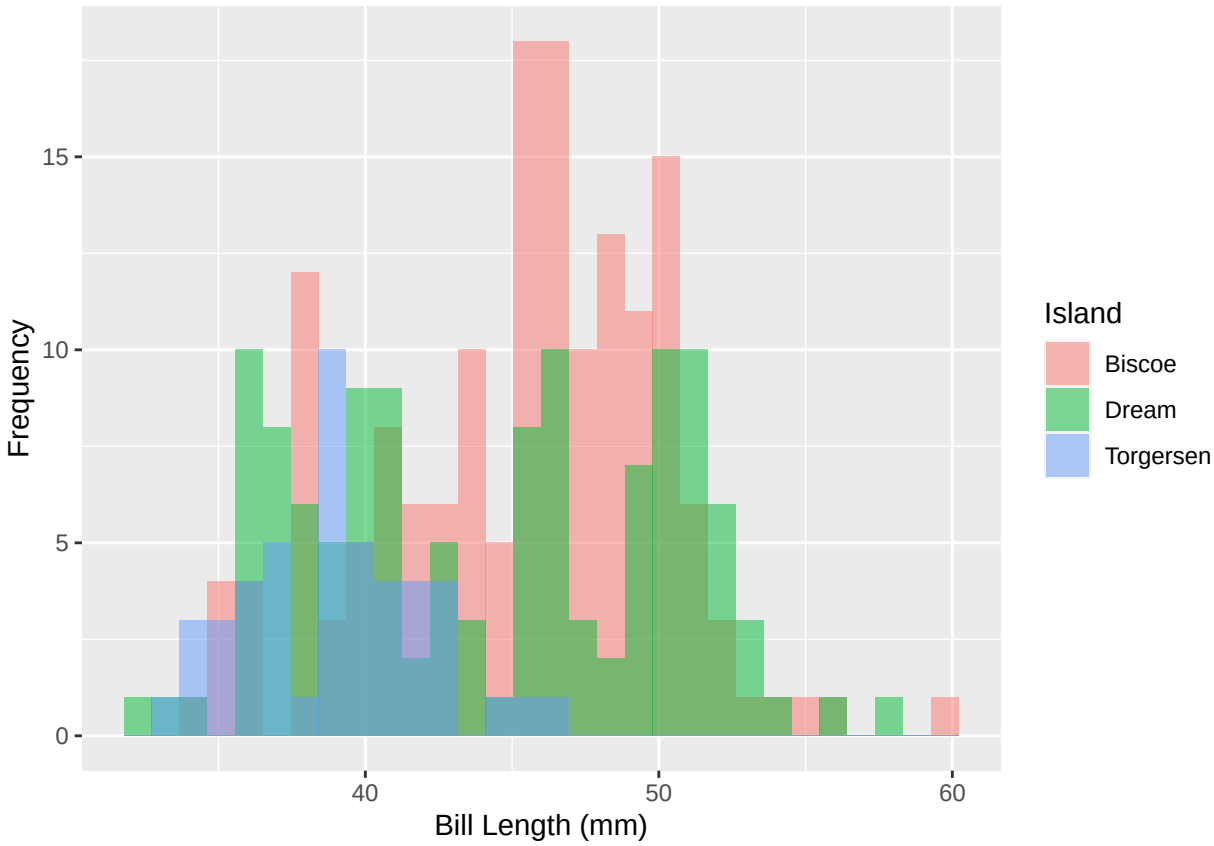


```
# OR

ggplot(penguins, aes(bill_length_mm, fill = island)) +
  geom_density(alpha = 0.5) +
  labs(x = "Bill Length (mm)",
       y = "Density",
       fill = "Island")
```




```
# OR

ggplot(penguins, aes(bill_length_mm, fill = island)) +
  geom_histogram(alpha = 0.5, position = "identity") +
  labs(x = "Bill Length (mm)",
       y = "Frequency",
       fill = "Island")
```

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



4. Write the correct pair of statistical hypotheses for our question. (2 points)

   Null: no difference in the mean bill length between penguins on different islands
   Alternative: true difference in the mean bill length between penguins on different islands

5. Run the appropriate statistical analysis for our question. (2 points)

```
aov_model <- aov(data = penguins, bill_length_mm ~ island)
summary(aov_model)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## island        2   1417   708.6   27.47 9.21e-12 ***
## Residuals   330   8512    25.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6. Interpret the results of this test. (3 points)

   - is there a significant difference?
   - what does that significant difference mean?
   - should we reject the null hypothesis?

   Answer: yes, p = 9.21e-12, which is smaller than 0.05, reject null. penguins on different islands have significantly different beak lengths

7. Should we run pairwise comparisons? If no, explain why not. If yes, do so and interpret the results. (3 points)

   Answer: yes; significant differences between all pairs except Dream-Biscoe because p > 0.05

```
TukeyHSD(aov_model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = bill_length_mm ~ island, data = penguins)
##
## $island
##                        diff       lwr        upr     p adj
## Dream-Biscoe     -1.026515 -2.454611  0.4015809 0.2096455
## Torgersen-Biscoe -6.210168 -8.189815 -4.2305220 0.0000000
## Torgersen-Dream  -5.183653 -7.234077 -3.1332293 0.0000000
```

Problem Set 2 (20 points)

Question: Is there a significant relationship between bill length and bill depth for penguins on Biscoe Island? For this problem set, we are going to use data from Biscoe island only.

1. Create a new data frame called `biscoe` that includes only penguins from Biscoe island. This new data frame should have 163 rows. (1 point)

```
biscoe <- penguins %>%
  filter(island == "Biscoe")
```

You will want to use the `biscoe` data set for the rest of Problem Set #2.

This is a scenario where there is no independent and no dependent variable. Go ahead and **put bill length on the x-axis and bill depth on the y-axis.**

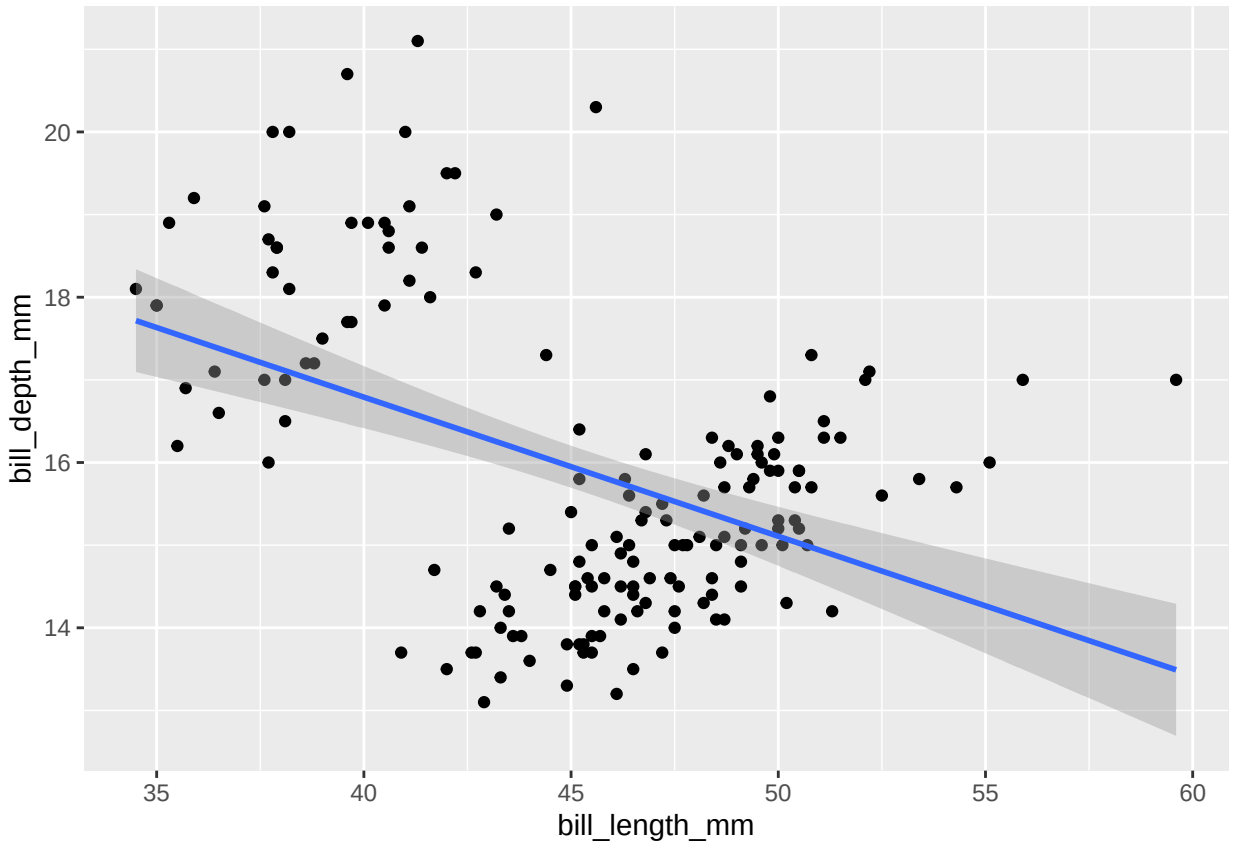
For now, ignore species. We will address that later in the problem set.

2. Plot the relationship between bill length and bill depth using the appropriate plot type. (2 points)

    - Be sure to add a line of best fit using the `geom_smooth` function—and make sure it is a straight line (no wiggles, which the default will produce).
    - Ensure that the plot has clear labels on the axes (follow the Structure & Guidelines).
    - Remember, we are ignoring species for now.

```
ggplot(biscoe, aes(bill_length_mm, bill_depth_mm)) +
  geom_point() +
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



3. Calculate the correlation coefficient, `r`. What does this value tell us about the relationship (positive, negative, no relationship)? (2 points)

   Answer: negative relationship

```
r <- cor(biscoe$bill_length_mm, biscoe$bill_depth_mm)
```

4. Calculate the `r^2` value. How much variation is explained by the line of best fit? Remember, this number is typically expressed as a percent (x 100). (2 points)

   Answer: About 20% variation explained

```
r^2
```

```
## [1] 0.1977277
```

5. Let's see if there is a significant relationship between bill length and bill depth. Perform the correct statistical analysis (1 point) and interpret the results. (4 points total)

   - What is the equation of the line of best fit? (1 point)
   - What is the p-value? (1 point)
   - Is there a significant relationship? (1 point)

   *Answer: bill_depth = -0.168*bill_length + 23.5; p = 2.74e-9, highly significant*

```
summary(lm(data = biscoe, bill_depth_mm ~ bill_length_mm))
```

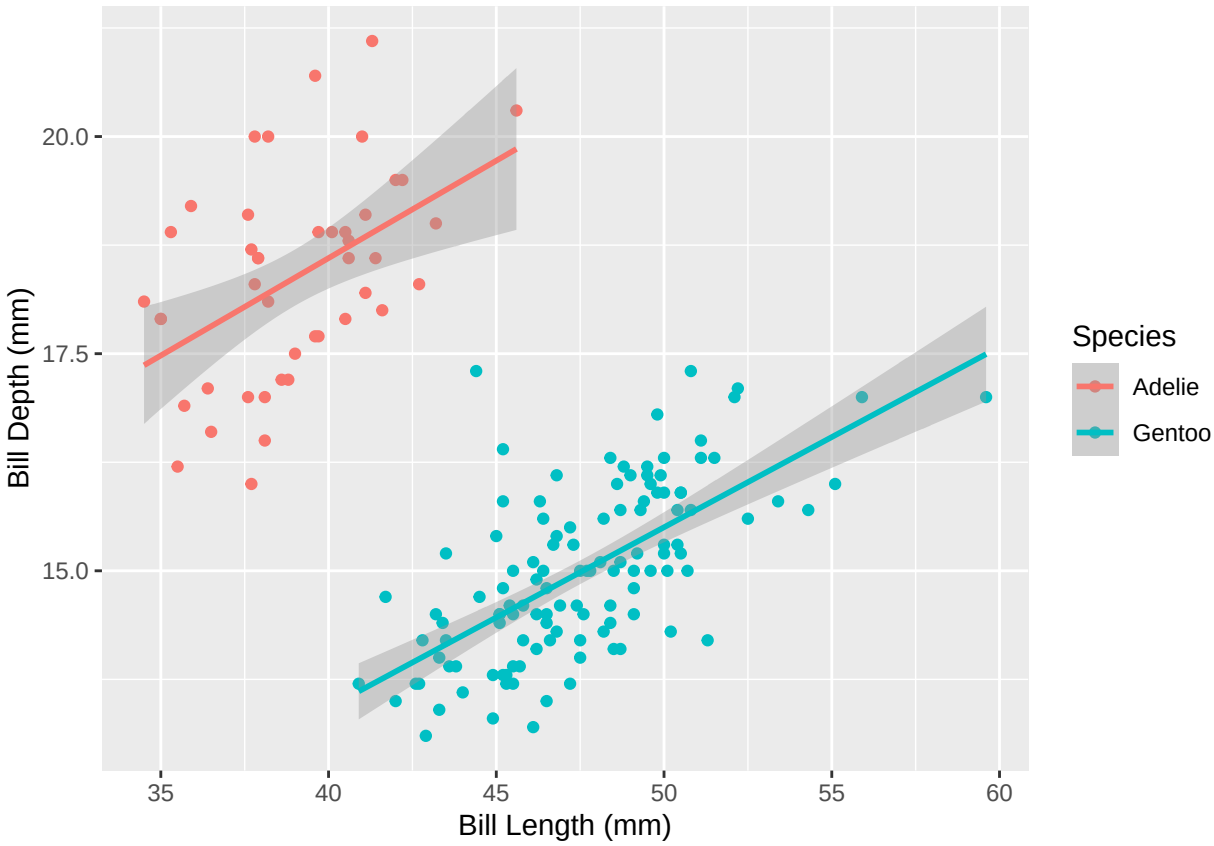
```
##
## Call:
## lm(formula = bill_depth_mm ~ bill_length_mm, data = biscoe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2027 -1.2536 -0.1135  1.0735  4.5279
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    23.52509    1.21614  19.344  < 2e-16 ***
## bill_length_mm -0.16835    0.02673  -6.299 2.74e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.642 on 161 degrees of freedom
## Multiple R-squared:  0.1977, Adjusted R-squared:  0.1927
## F-statistic: 39.68 on 1 and 161 DF,  p-value: 2.738e-09
```

When we look at the plot of the data, it looks like there might be two different groups in the data.

6. Let's make the color of the points and the linear models differ by species on Biscoe Island. Be sure to adjust *all* labels on the plot accordingly. (2 points)

```
ggplot(biscoe, aes(bill_length_mm, bill_depth_mm, color = species)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Bill Length (mm)",
       y = "Bill Depth (mm)",
       color = "Species")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



7. Run the appropriate statistical test, adding species into the model. (2 points)

```
summary(lm(data = biscoe, bill_depth_mm ~ bill_length_mm * species))
```

```
##
## Call:
## lm(formula = bill_depth_mm ~ bill_length_mm * species, data = biscoe)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.08441 -0.64479 -0.02957  0.46954  2.96109
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     9.62634    2.02455   4.755 4.42e-06 ***
## bill_length_mm                  0.22435    0.05184   4.328 2.66e-05 ***
## speciesGentoo                  -4.50539    2.34915  -1.918   0.0569 .
## bill_length_mm:speciesGentoo -0.01674    0.05755  -0.291   0.7715
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8434 on 159 degrees of freedom
## Multiple R-squared:  0.791,  Adjusted R-squared:  0.7871
## F-statistic: 200.6 on 3 and 159 DF,  p-value: < 2.2e-16
```

8. Interpret the results of the test above. Is species significant? Is there a significant interaction between bill length and species? (2 points)

Answer: neither species nor interaction is significant

9. Write 2-3 sentences discussing if and/or how adding species into our linear model changes our interpretation of the data. Did the type of relationship change? Do the two linear models tell us different things? (3 points)

   Answer: negative to positive; models still tell us similar things in terms of significance, though

Problem Set 3 (20 points)

Question: Is there a difference in the average *body mass* between penguins with long flippers and penguins with short flippers? In order to address this question, our first step is to create a new column in our data frame that tells us whether each penguin has a long flipper or a short flipper.

- flippers are considered "long" if they are at least 200 mm in length (>= 200)
- flippers are considered "short" if they are less than 200 mm (< 200)

1. Using the `mutate()` and `if_else()` functions, create a new column called `long_or_short` with the correct term ("long" or "short") for each flipper length. Be sure to save this in a new data frame called `flippers`. (3 points)

```
flippers <- penguins %>%
  mutate(long_or_short = if_else(flipper_length_mm >= 200, "long", "short"))
```

We will be using the `flippers` data frame for the rest of this problem set! It has the `long_or_short` column that we will be referencing.

2. Now let's summarize our data using this new column. Calculate the mean and standard deviation *body mass* for penguins with long flippers and penguins with short flippers. (2 points)

   Take a few seconds to think this through before you start coding. We want the values for *body mass*. We want our groups determined by the values in the `long_or_short` column that we just created.

```
flippers %>%
  group_by(long_or_short) %>%
  summarise(mean_mass = mean(body_mass_g),
            sd_mass = sd(body_mass_g))
```

```
## # A tibble: 2 x 3
##   long_or_short mean_mass sd_mass
##   <chr>             <dbl>   <dbl>
## 1 long              4894.    621.
## 2 short             3657.    422.
```

Ok, we have summarized the body mass data for our two groups, and it looks like there might be a real difference in the body masses between the long group and the short group.

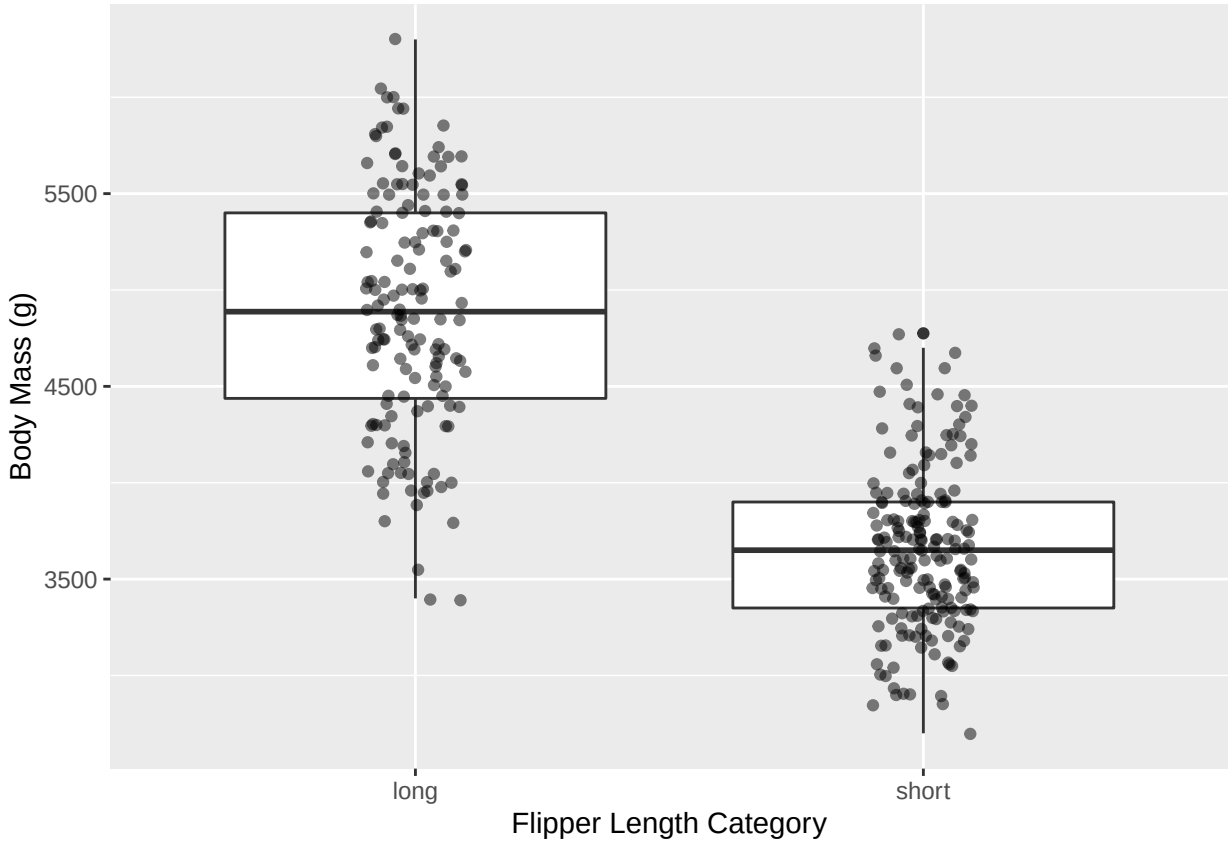
3. Determine which variable is dependent and which is independent. Also determine if each variable is continuous or categorical. (4 points)

- **body mass**: dependent, continuous
- **flipper group (`long_or_short`)**: independent, categorical

Let's plot the body mass data for the two groups.

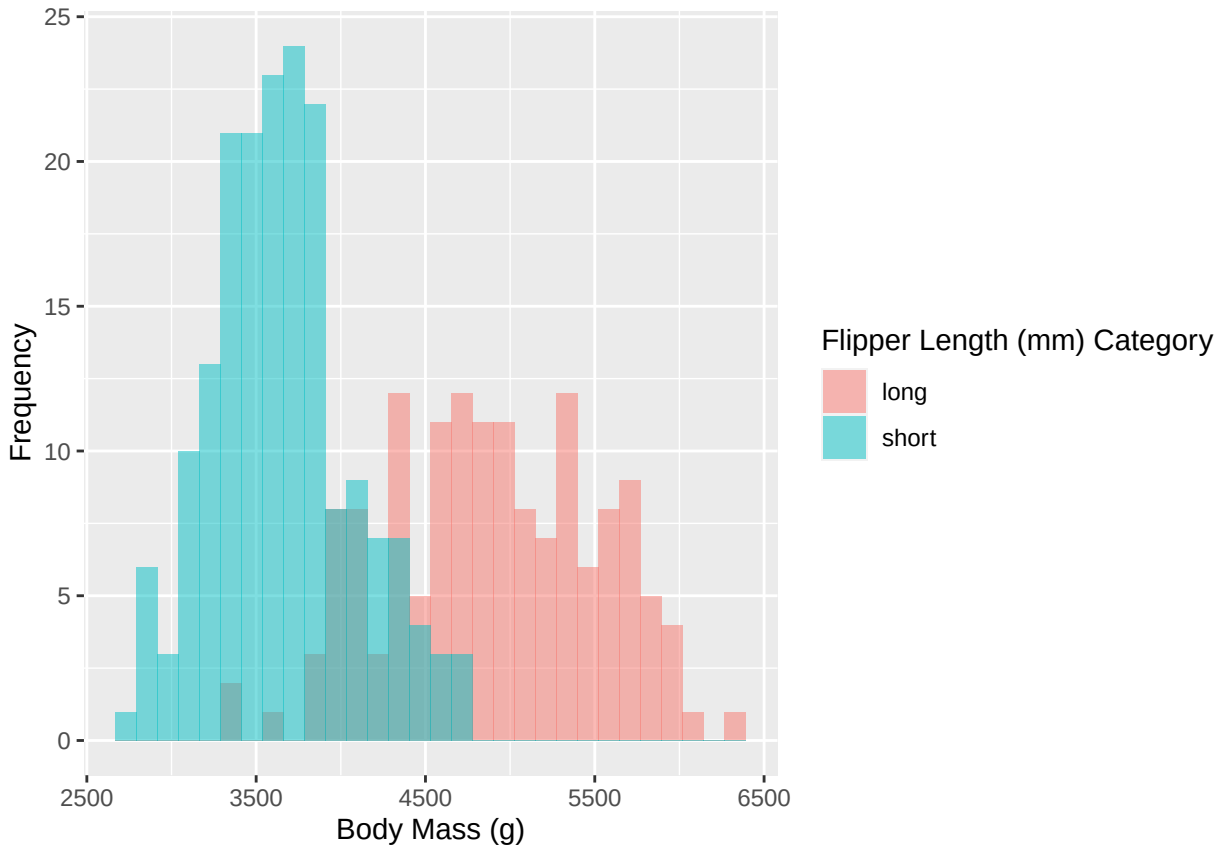
4. Choose an appropriate plot for data with one continuous variable and one categorical variable (there are a few options). Be sure to adjust the x- and y-axis labels appropriately. (2 points)

```
ggplot(flippers, aes(long_or_short, body_mass_g)) +
  geom_boxplot() +
  geom_jitter(width = 0.1, alpha = 0.5) +
  labs(x = "Flipper Length Category",
       y = "Body Mass (g)")
```

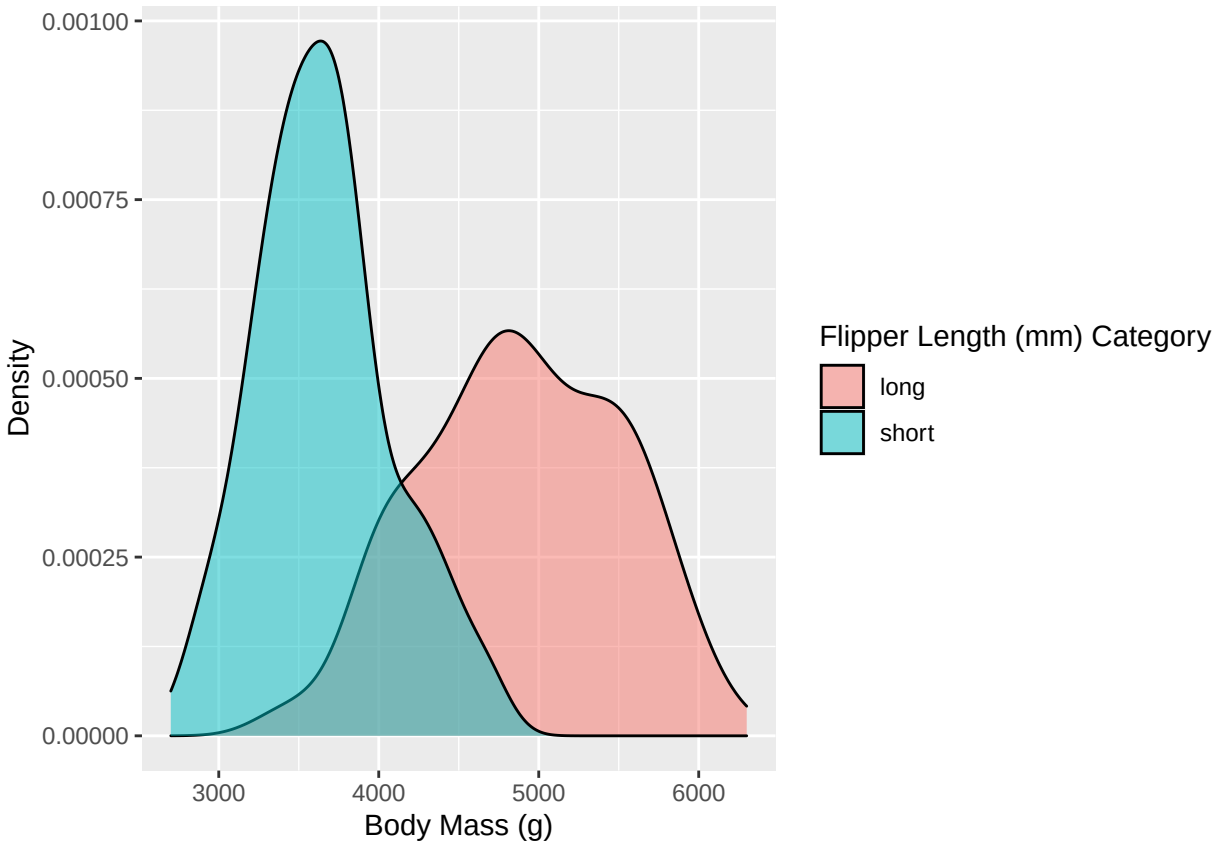



```
# OR

ggplot(flippers, aes(body_mass_g, fill = long_or_short)) +
  geom_histogram(alpha = 0.5, position = "identity") +
  labs(x = "Body Mass (g)",
       y = "Frequency",
       fill = "Flipper Length (mm) Category")
```

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




```
# OR

ggplot(flippers, aes(body_mass_g, fill = long_or_short)) +
  geom_density(alpha = 0.5) +
  labs(x = "Body Mass (g)",
       y = "Density",
       fill = "Flipper Length (mm) Category")
```



5. Write the pair of statistical hypotheses for our question. (2 points)

   Null: no difference in the mean body mass between penguins with short flippers and long flippers
   Alternative: true difference in the mean body mass between penguins with short flippers and long flippers

6. Perform the appropriate analysis to compare the body mass means of our two groups: long and short. (2 points)

```
t.test(data = flippers, body_mass_g ~ long_or_short)
```

```
##
##  Welch Two Sample t-test
##
## data:  body_mass_g by long_or_short
## t = 20.696, df = 248.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group long and group short is not equal to 0
## 95 percent confidence interval:
##  1118.967 1354.344
## sample estimates:
##  mean in group long mean in group short
##            4894.088            3657.432
```

7. Interpret the results of this test. (3 points)

   - is there a significant difference?
   - what does that significant difference mean?

- should we reject the null hypothesis?

Answer: yes, there is a significant difference in the means of body mass between penguins with long flippers and short flippers (p = 2.2e-16); reject null

8. Should we run pairwise comparisons? If no, explain why not. If yes, do so and interpret the results. (2 points)

   Answer: no, t-test is comparing just one pair

The End!

Great work, and thanks for a wonderful semester!