Problem set 3: Regression

David Segovia

10/11/22


```
library(tidyverse)
library(broom)
library(modelsummary)


# Load penguins data
setwd("~/Downloads/PA 528 Public Program Evaluation/Problem Sets/PS3")
penguins <- read_csv("penguins.csv")
```

Task 1: Penguins

Between 2007 and 2009, researchers collected data on penguins in three islands in the Palmer Archipelago in Antarctica: Biscoe, Dream, and Torgersen. The `penguins` dataset has data for 342 penguins from 3 different species: Chinstrap, Gentoo, and Adélie. It includes the following variables:

- `species`: The penguin's species (Chinstrap, Gentoo, and Adélie)
- `island`: The island where the penguin lives (Biscoe, Dream, and Torgersen)
- `bill_length_mm`: The length of the penguin's bill, in millimeters (distance from the penguin's face to the tip of the bill)
- `bill_depth_mm`: The depth of the penguin's bill, in millimeters (height of the bill; distance from the bottom of the bill to the top of the bill)
- `flipper_length_mm`: The length of the penguin's flippers, in millimeters
- `body_mass_g`: The weight of the penguin, in grams
- `sex`: The sex of the penguin
- `year`: The year the observation was made

Exploratory analysis

How related are penguin weight and bill depth?

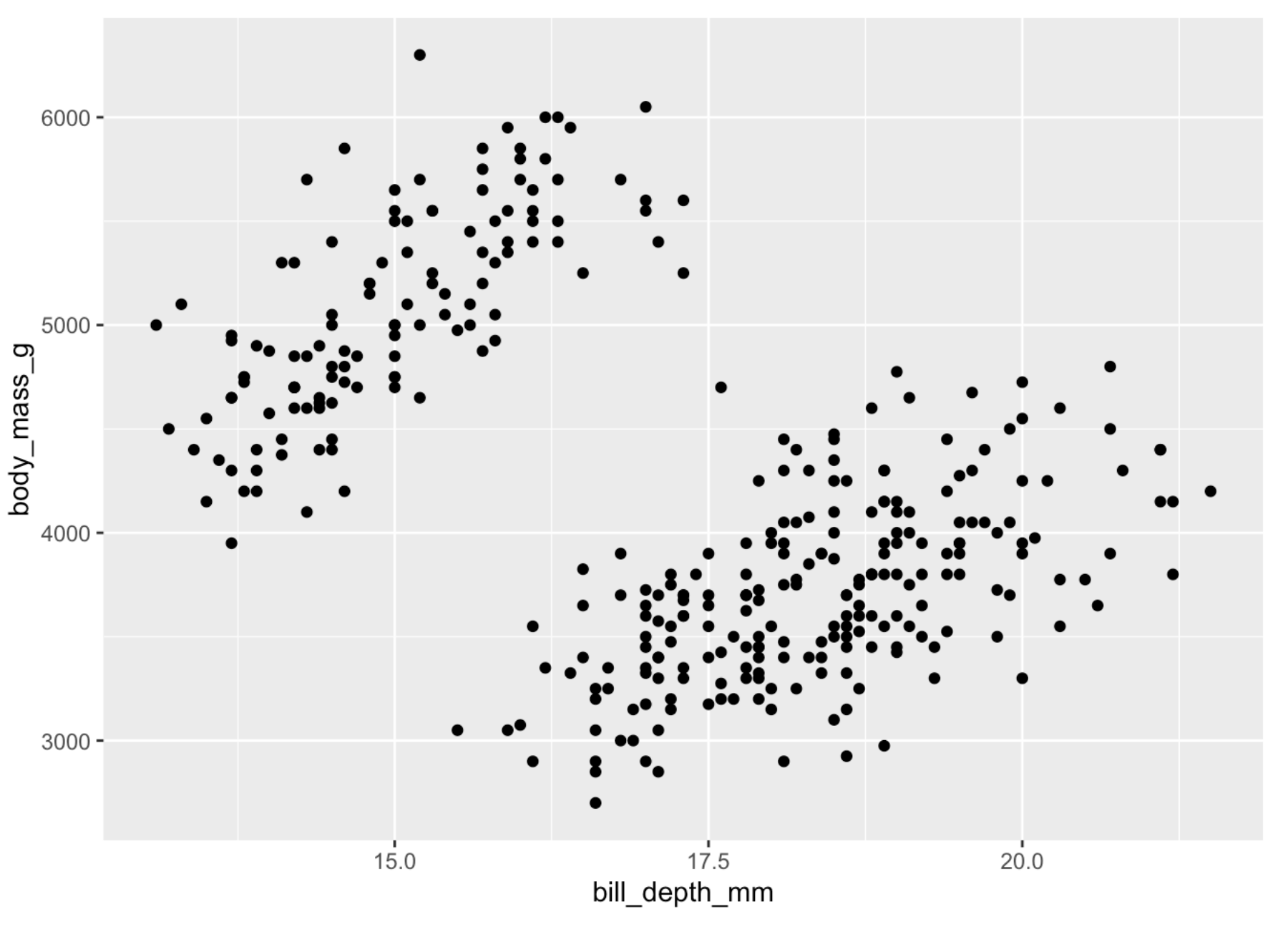
Penguin weight and bill depth seem to have a moderately strong, negative correlation.

```
# It's not possible to calculate correlations when there is missing data.
# The `use = "complete.obs"` argument here tells R to ignore any
# rows where either mortality_rate or pct_low_access_pop is missing

cor(penguins$bill_depth_mm, penguins$body_mass_g,
    use = "complete.obs")
```

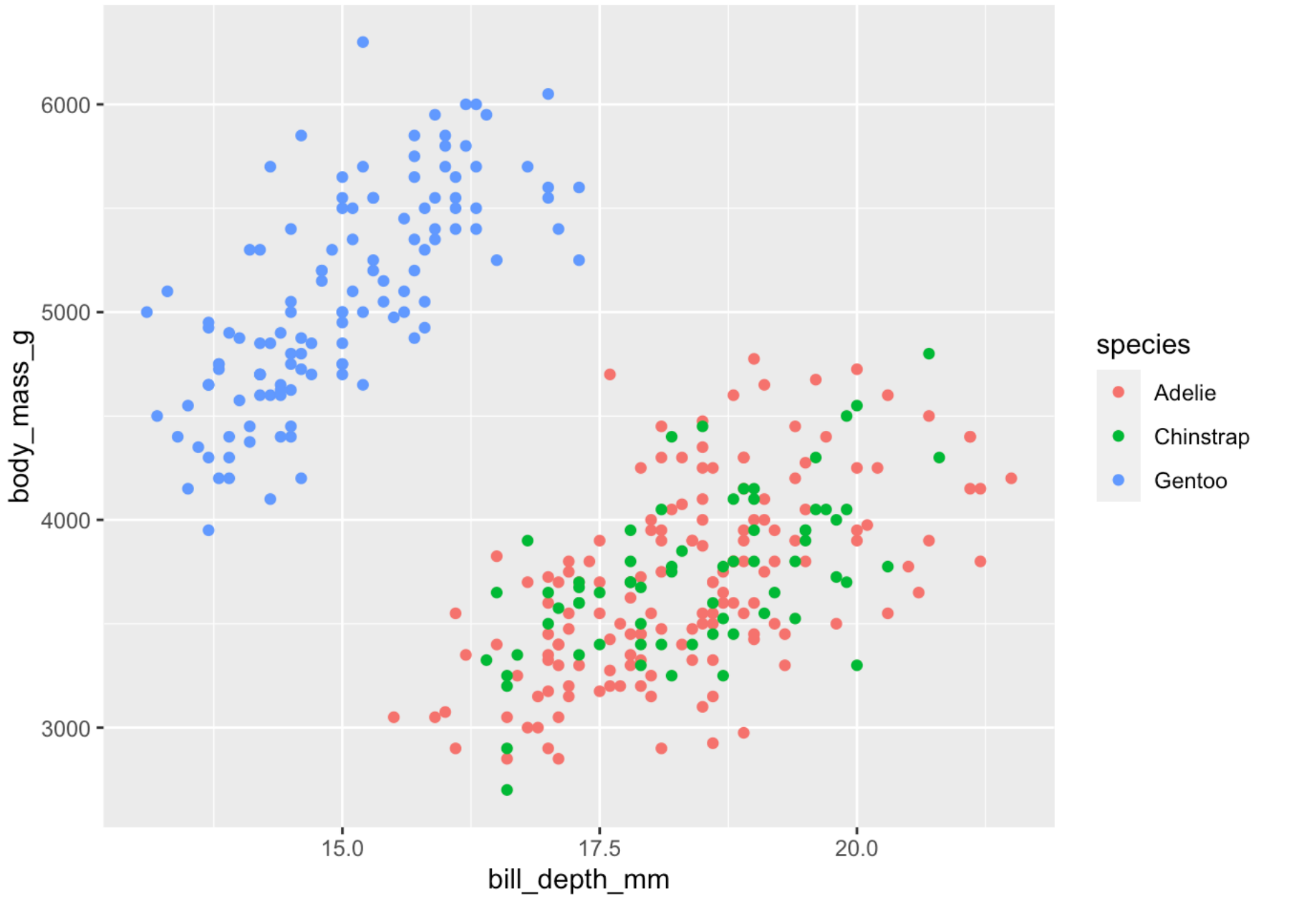
```
## [1] -0.4719156
```

```
ggplot(data = penguins,
       aes(x = bill_depth_mm, y = body_mass_g)) +
  geom_point()
```



Make a new plot that colors these points by species. What can you tell about the relationship between bill depth and penguin weight?

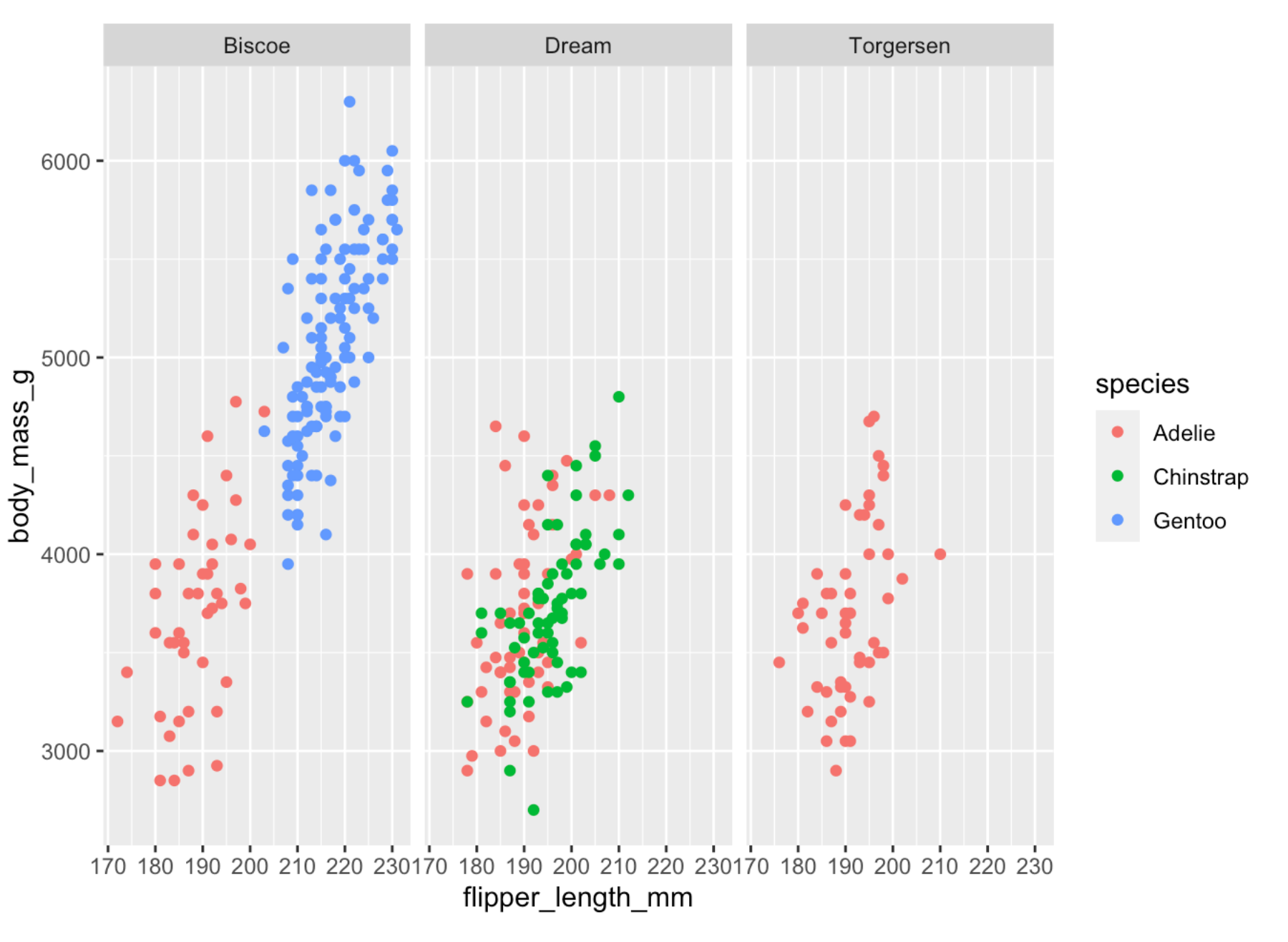
```
ggplot(data = penguins,
       aes(x = bill_depth_mm, y = body_mass_g, color = species)) +
  geom_point()
```



For all penguin species "Adelie", "Chinstrap", and "Gentoo", the relationship between bill depth and penguin weight is positively correlated. As bill depth increases, so does weight. This relationship contrasts from the correlation earlier.

What is the relationship between flipper length and body mass? Make another plot with `flipper_length_mm` on the x-axis, `body_mass_g` on the y-axis, and points colored by `species`. Facet the plot by island (`island`)

```
ggplot(data = penguins,
       aes(x = flipper_length_mm, y = body_mass_g, color = species)) +
  geom_point() +
  facet_wrap(vars(island))
```



Tell a story about the relationship between flipper length and weight in these three penguin species.

In Biscoe Island, Gentoo and Adelie penguins exist. However, Gentoo species have a stronger relationship between body mass and flipper length, as Gentoo species are bigger in size and have bigger flippers. Dream Island seems to have Adelie and Chinstrap species. These two also have a positive relationship between body mass and flipper length. However, only Adelie penguins exist in Torgersen, and they have a positive relationship as well between flipper length and body mass.

Tell a story about the distribution of penguins across the three islands.

In Biscoe island, Gentoo and Adelie penguin species exist. In Dream island, Adelie and Chinstrap penguins exist. In Torgersen island, only Adelie penguins exist. Adelie penguins exist across all 3 islands.

Models

Predicting weight with bill depth

Does bill depth predict penguin weight?

Yes, it does. It is statistically significant (p<.05) and the relationship is negative.

```
model_depth_weight <- lm(body_mass_g ~ bill_depth_mm,
                         data = penguins)
summary(model_depth_weight)
```

```
##
## Call:
## lm(formula = body_mass_g ~ bill_depth_mm, data = penguins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1607.38  -510.10   -66.96   462.43  1819.28
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7488.65     335.22   22.34   <2e-16 ***
## bill_depth_mm  -191.64      19.42   -9.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 708.1 on 340 degrees of freedom
## Multiple R-squared:  0.2227, Adjusted R-squared:  0.2204
## F-statistic: 97.41 on 1 and 340 DF,  p-value: < 2.2e-16
```

```
tidy(model_depth_weight, conf.int = TRUE)
```

```
## # A tibble: 2 x 7
##   term          estimate std.error statistic  p.value conf.low conf.high
##   <chr>            <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)      7489.     335.      22.3  1.13e-68    6829.     8148.
## 2 bill_depth_mm    -192.      19.4     -9.87 2.28e-20    -230.     -153.
```

```
glance(model_depth_weight)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.223         0.220  708.      97.4 2.28e-20     1 -2729. 5463. 5475.
## # … with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

INTERPRET THE COEFFICIENTS AND RESULTS HERE. What happens as bills get taller? Is the association statistically significant? How confident are you about these results? (Hint: look at the R^2)

As bills get taller, weight decreases.For one mm increase in bill depth, weight decreases by 191.64 grams. This is statistically significant at the p<.001 level. We are 22.27% confident that this model accurately predicts a penguin's weight.

Predicting weight with bill depth and flipper length

RUN A MODEL that predicts weight with bill depth and flipper length (i.e. body_mass_g ~ bill_depth_mm + flipper_length_mm)

```
model_depth_weight2 <- lm(body_mass_g ~ bill_depth_mm + flipper_length_mm,
                          data = penguins)
summary(model_depth_weight2)
```

```
##
## Call:
## lm(formula = body_mass_g ~ bill_depth_mm + flipper_length_mm,
##     data = penguins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1029.78  -271.45   -23.58   245.15  1275.97
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -6541.907    540.751 -12.098   <2e-16 ***
## bill_depth_mm        22.634     13.280   1.704   0.0892 .
## flipper_length_mm    51.541      1.865  27.635   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 393.2 on 339 degrees of freedom
## Multiple R-squared:  0.761,  Adjusted R-squared:  0.7596
## F-statistic: 539.8 on 2 and 339 DF,  p-value: < 2.2e-16
```

```
tidy(model_depth_weight2, conf.int = TRUE)
```

```
## # A tibble: 3 x 7
##   term              estimate std.error statistic  p.value conf.low conf.high
##   <chr>                <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)        -6542.     541.       -12.1  2.99e-28 -7606.     -5478.
## 2 bill_depth_mm         22.6     13.3        1.70 8.92e- 2    -3.49      48.8
## 3 flipper_length_mm     51.5      1.87      27.6  7.72e-89    47.9       55.2
```

```
glance(model_depth_weight2)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic   p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>     <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.761         0.760  393.      540. 4.23e-106     2 -2527. 5062. 5077.
## # … with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

INTERPRET THESE RESULTS. Did the size of the bill depth coefficient change after controlling for flipper length?

Yes. The size of bill depth went from negative to positive and smaller, and bill depth is no longer a predictor variable that predicts penguin weight when controlling for flipper length. In fact, flipper length became a significant predictor variable at the p<.001 level.

Predicting weight with bill depth, flipper length, and species

RUN A MODEL that predicts weight with bill depth, flipper length, and species.

```
model_depth_weight3 <- lm(body_mass_g ~ bill_depth_mm + flipper_length_mm + species,
                          data = penguins)
summary(model_depth_weight3)
```

```
##
## Call:
## lm(formula = body_mass_g ~ bill_depth_mm + flipper_length_mm +
##     species, data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -900.21 -237.93  -39.51  228.11 1086.25
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -4526.887    516.931  -8.757  < 2e-16 ***
## bill_depth_mm       182.364     18.358   9.934  < 2e-16 ***
## flipper_length_mm    25.700      3.098   8.295 2.63e-15 ***
## speciesChinstrap   -131.968     51.400  -2.567   0.0107 *
## speciesGentoo      1288.968    132.774   9.708  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 330.8 on 337 degrees of freedom
## Multiple R-squared:  0.8319, Adjusted R-squared:  0.8299
## F-statistic: 416.9 on 4 and 337 DF,  p-value: < 2.2e-16
```

```
tidy(model_depth_weight3, conf.int = TRUE)
```

```
## # A tibble: 5 x 7
##   term              estimate std.error statistic  p.value conf.low conf.high
##   <chr>                <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)        -4527.     517.       -8.76 9.87e-17  -5544.    -3510.
## 2 bill_depth_mm        182.      18.4       9.93 1.45e-20    146.      218.
## 3 flipper_length_mm     25.7      3.10      8.30 2.63e-15     19.6      31.8
## 4 speciesChinstrap    -132.      51.4      -2.57 1.07e- 2   -233.      -30.9
## 5 speciesGentoo       1289.     133.        9.71 8.28e-20   1028.     1550.
```

```
glance(model_depth_weight3)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic   p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>     <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.832         0.830  331.      417. 4.66e-129     4 -2467. 4946. 4969.
## # … with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

INTERPRET THESE RESULTS. What do the species coefficients mean? Did the bill depth coefficient change after controlling for both flipper length and species?

The species coefficients means that after controlling for bill depth and flipper length, species is a significant predictor variable that predicts penguin weight. The base group here is Adelie, so Adelie penguins weigh 131 grams more than Chinstrap penguins but 1288 grams less than Gentoo penguins. All of these variables are significant at the .05 level *. Bill depth coefficient, when controlling for species and flipper length, also became a significant predictor variable and is positive.

This time, we are 83.2% confident that this model accurately predicts a penguin's weight.

All models at the same time

```
# Right now there's only one model here. Add the others from above (whatever you
# called them) like so:
# modelsummary(list(model_depth_weight, some_other_model, yet_another_model, etc))
 modelsummary(list(model_depth_weight,model_depth_weight2,model_depth_weight3))
```

	Model 1	Model 2	Model 3
(Intercept)	7488.652	−6541.907	−4526.887
	(335.218)	(540.751)	(516.931)
bill_depth_mm	−191.643	22.634	182.364
	(19.417)	(13.280)	(18.358)
flipper_length_mm		51.541	25.700
		(1.865)	(3.098)
speciesChinstrap			−131.968
			(51.400)
speciesGentoo			1288.968
			(132.774)
Num.Obs.	342	342	342
R2	0.223	0.761	0.832
R2 Adj.	0.220	0.760	0.830
AIC	5463.3	5061.9	4945.7
BIC	5474.8	5077.3	4968.7
Log.Lik.	−2728.667	−2526.968	−2466.846
F	97.414	539.824	416.867