HW 4

Enter your name and EID here: Joseph Hendrix | jlh7459

You will submit this homework assignment as a pdf file on Gradescope.

For all questions, include the R commands/functions that you used to find your answer (show R chunk). Answers without supporting code will not receive credit. Write full sentences to describe your findings.

Question 1: (2 pts)

All subsequent code will be done using `dplyr`, so we need to load this package. We also want to look at the `penguins` dataset which is inside the `palmerpenguins` package:

```
# Call dplyr and ggplot2 packages within tidyverse
library(tidyverse)

# Paste and run the following uncommented code into your console:
# install.packages("palmerpenguins")

# Save the data as a dataframe
penguins <- as.data.frame(palmerpenguins::penguins)
```

Using a `dplyr` function, pick all the rows/observaions in the `penguins` dataset from the year 2007 and save the result as a new object called `penguins_2007`. Compare the number of observations/rows in the original `penguins` dataset with your new `penguins_2007` dataset.

```
# filtering penguins data set
penguins_2007 <- penguins %>% filter(year == 2007)
penguins
```

```
##    species    island bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
## 1   Adelie Torgersen           39.1          18.7               181        3750
## 2   Adelie Torgersen           39.5          17.4               186        3800
## 3   Adelie Torgersen           40.3          18.0               195        3250
## 4   Adelie Torgersen             NA            NA                NA          NA
## 5   Adelie Torgersen           36.7          19.3               193        3450
## 6   Adelie Torgersen           39.3          20.6               190        3650
## 7   Adelie Torgersen           38.9          17.8               181        3625
## 8   Adelie Torgersen           39.2          19.6               195        4675
## 9   Adelie Torgersen           34.1          18.1               193        3475
## 10  Adelie Torgersen           42.0          20.2               190        4250
## 11  Adelie Torgersen           37.8          17.1               186        3300
## 12  Adelie Torgersen           37.8          17.3               180        3700
##       sex year
## 1    male 2007
## 2  female 2007
## 3  female 2007
## 4    <NA> 2007
## 5  female 2007
## 6    male 2007
## 7  female 2007
## 8    male 2007
## 9    <NA> 2007
## 10   <NA> 2007
## 11   <NA> 2007
## 12   <NA> 2007
##  [ reached 'max' / getOption("max.print") -- omitted 332 rows ]
```

```
penguins_2007
```

```
##    species    island bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
## 1   Adelie Torgersen           39.1          18.7               181        3750
## 2   Adelie Torgersen           39.5          17.4               186        3800
## 3   Adelie Torgersen           40.3          18.0               195        3250
## 4   Adelie Torgersen             NA            NA                NA          NA
## 5   Adelie Torgersen           36.7          19.3               193        3450
## 6   Adelie Torgersen           39.3          20.6               190        3650
## 7   Adelie Torgersen           38.9          17.8               181        3625
## 8   Adelie Torgersen           39.2          19.6               195        4675
## 9   Adelie Torgersen           34.1          18.1               193        3475
## 10  Adelie Torgersen           42.0          20.2               190        4250
## 11  Adelie Torgersen           37.8          17.1               186        3300
## 12  Adelie Torgersen           37.8          17.3               180        3700
##       sex year
## 1    male 2007
## 2  female 2007
## 3  female 2007
## 4    <NA> 2007
## 5  female 2007
## 6    male 2007
## 7  female 2007
## 8    male 2007
## 9    <NA> 2007
## 10   <NA> 2007
## 11   <NA> 2007
## 12   <NA> 2007
##  [ reached 'max' / getOption("max.print") -- omitted 98 rows ]
```

The full penguins dataset has 344 observations, 110 of which are from 2007

Question 2: (2 pts)

Using `dplyr` functions on `penguins_2007`, report the number of observations for each species-island combination (note that you'll need to `group_by`). Which species appears on all three islands?

```
# pulling species-island count
penguins_2007 %>%
  group_by(species, island) %>%
  summarize(n())
```

```
## # A tibble: 5 × 3
## # Groups:   species [3]
##   species   island    `n()`
##   <fct>     <fct>     <int>
## 1 Adelie    Biscoe       10
## 2 Adelie    Dream        20
## 3 Adelie    Torgersen    20
## 4 Chinstrap Dream        26
## 5 Gentoo    Biscoe       34
```

Adelie penguins appear on all three islands.

Question 3: (2 pts)

Using `dplyr` functions on `penguins_2007`, create a new variable that contains the ratio of `bill_length_mm` to `bill_depth_mm` (call it `bill_ratio`). *Once you checked that your variable is created correctly*, overwrite `penguins_2007` so it contains this new variable.

```
# your code goes below (make sure to edit comment)
penguins_2007$bill_length_mm / penguins_2007$bill_depth_mm    # verifying
```

```
##   [1] 2.090909 2.270115 2.238889       NA 1.901554 1.907767 2.185393 2.000000
##   [9] 1.883978 2.079208 2.210526 2.184971 2.335227 1.820755 1.639810 2.056180
##  [17] 2.036842 2.053140 1.869565 2.139535 2.065574 2.016043 1.869792 2.110497
##  [25] 2.255814 1.867725 2.182796 2.262570 2.037634 2.142857 2.365269 2.055249
##  [33] 2.219101 2.164021 2.141176 1.857820 1.940000 2.281081 1.948187 2.083770
##  [41] 2.027778 2.217391 1.945946 2.238579 2.189349 2.106383 2.163158 1.984127
##  [49] 2.011173 1.995283 3.492424 3.067485 3.453901 3.289474 3.282759 3.444444
##  [57] 3.109589 3.052288 3.231343 3.038961 2.985401 3.043478 3.321168 3.315068
##  [65] 3.136986 3.140127 3.111111 3.236842 3.186207 3.225166 3.510490 3.110345
##  [73] 3.206897 2.930380 3.274809 3.052980 3.111888 3.186667 3.370629 3.267974
##  [81] 3.091503 3.014085 3.110345 3.505882 2.597765 2.564103 2.671875 2.427807
##  [89] 2.661616 2.539326 2.532967 2.818681 2.433862 2.577889 2.617978 2.546798
##  [97] 2.716763 2.872928 2.684211 2.576531
##  [ reached getOption("max.print") -- omitted 10 entries ]
```

```
penguins_2007 %>%
  mutate(bill_ratio = (bill_length_mm / bill_depth_mm))    # adding to dataset
```

```
##    species    island bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
## 1   Adelie Torgersen           39.1          18.7               181        3750
## 2   Adelie Torgersen           39.5          17.4               186        3800
## 3   Adelie Torgersen           40.3          18.0               195        3250
## 4   Adelie Torgersen             NA            NA                NA          NA
## 5   Adelie Torgersen           36.7          19.3               193        3450
## 6   Adelie Torgersen           39.3          20.6               190        3650
## 7   Adelie Torgersen           38.9          17.8               181        3625
## 8   Adelie Torgersen           39.2          19.6               195        4675
## 9   Adelie Torgersen           34.1          18.1               193        3475
## 10  Adelie Torgersen           42.0          20.2               190        4250
## 11  Adelie Torgersen           37.8          17.1               186        3300
##       sex year bill_ratio
## 1    male 2007   2.090909
## 2  female 2007   2.270115
## 3  female 2007   2.238889
## 4    <NA> 2007         NA
## 5  female 2007   1.901554
## 6    male 2007   1.907767
## 7  female 2007   2.185393
## 8    male 2007   2.000000
## 9    <NA> 2007   1.883978
## 10   <NA> 2007   2.079208
## 11   <NA> 2007   2.210526
##  [ reached 'max' / getOption("max.print") -- omitted 99 rows ]
```

Are there any cases in the `penguins_2007` dataset for which the `bill_ratio` exceeds 3.5? If so, for which species of penguins is this true?

```
# select observations where bill ratio is greater than 3.5
penguins_2007 %>%
  mutate(bill_ratio = (bill_length_mm / bill_depth_mm)) %>%
  filter(bill_ratio > 3.5)
```

```
##   species island bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
## 1  Gentoo Biscoe           50.2          14.3               218        5700
## 2  Gentoo Biscoe           59.6          17.0               230        6050
##    sex year bill_ratio
## 1 male 2007   3.510490
## 2 male 2007   3.505882
```

The Gentoo penguin has a bill ratio greater than 3.5

Question 4: (2 pts)

Using `dplyr` functions on `penguins_2007`, find the three penguins with the smallest bill ratio for *each species*. Only display the information about `species`, `sex`, and `bill_ratio`. Does the same sex has the smallest bill ratio across species?

```
# your code goes below (make sure to edit comment)
penguins_2007 %>%
  mutate(bill_ratio = (bill_length_mm / bill_depth_mm), na.rm=T) %>%    # create bill_ratio
  select(species, sex, bill_ratio) %>%    # select relevant info
  group_by(species, sex) %>%
  summarize(min(bill_ratio, na.rm=T))    # select minimum bill ratio values
```

```
## # A tibble: 8 × 3
## # Groups:   species [3]
##   species   sex    `min(bill_ratio, na.rm = T)`
##   <fct>     <fct>                         <dbl>
## 1 Adelie    female                         1.87
## 2 Adelie    male                           1.64
## 3 Adelie    <NA>                           1.88
## 4 Chinstrap female                         2.43
## 5 Chinstrap male                           2.51
## 6 Gentoo    female                         2.99
## 7 Gentoo    male                           2.93
## 8 Gentoo    <NA>                           3.11
```

No! Males have smaller bill ratios among Adelie and Gentoo penguins but female Chinstraps have smaller bill ratios than male Chinstraps

---

Question 5: (2 pts)

Using `dplyr` functions on `penguins_2007`, calculate the mean and standard deviation of `bill_ratio` for each species. Drop NAs from `bill_ratio` for these computations (e.g., using the argument `na.rm = T`) so you have values for each species. Which species has the greatest mean `bill_ratio`?

```
penguins_2007 %>%
  mutate(bill_ratio = bill_length_mm / bill_depth_mm, na.rm = T) %>%    # create bill_ratio obs
  group_by(species) %>%
  summarize(mean(bill_ratio, na.rm=T), sd(bill_ratio, na.rm=T))    # calculate mean and sd of bi
ll ratio, excluding NA values
```

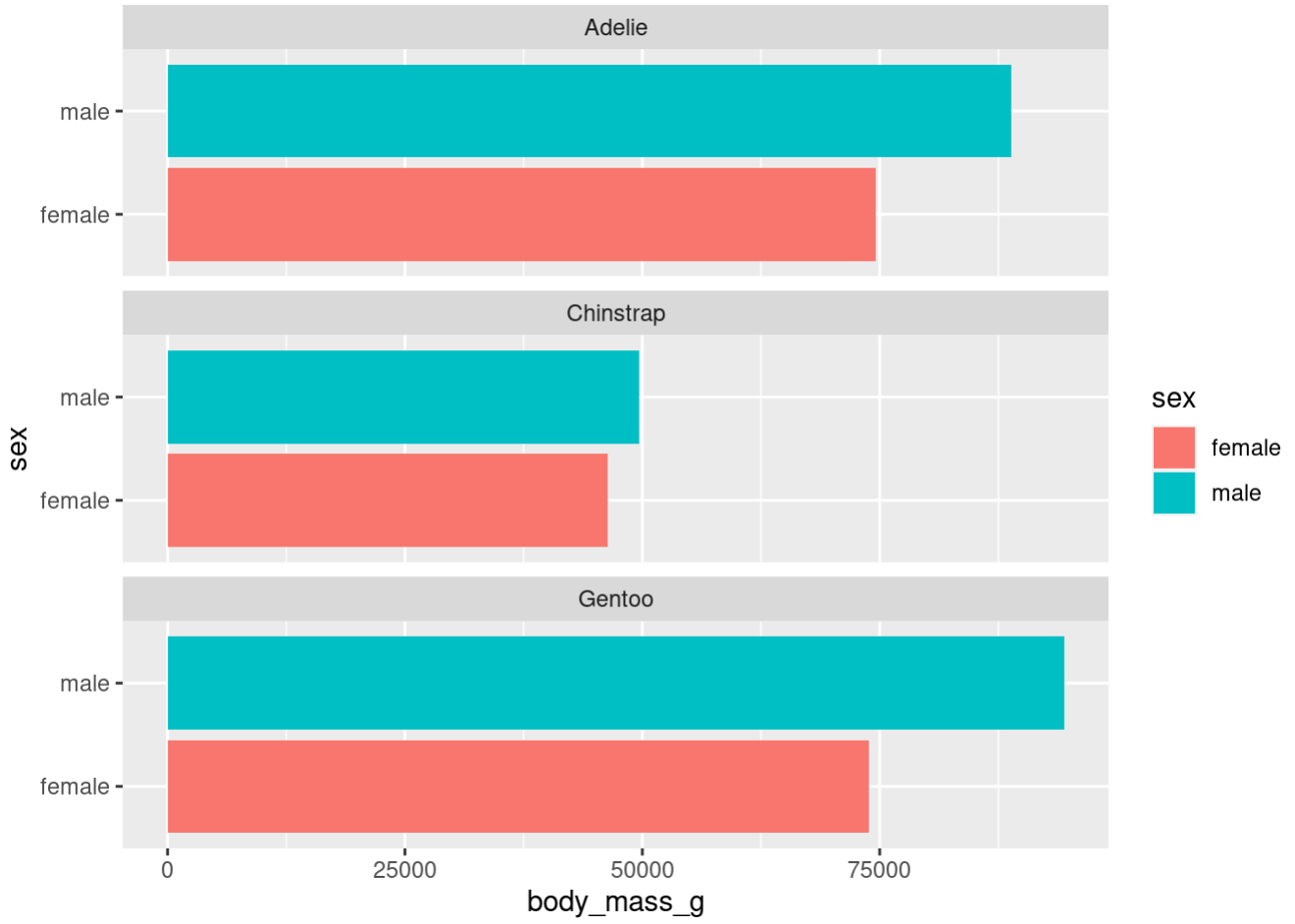
```
## # A tibble: 3 × 3
##   species   `mean(bill_ratio, na.rm = T)` `sd(bill_ratio, na.rm = T)`
##   <fct>                             <dbl>                       <dbl>
## 1 Adelie                             2.07                       0.152
## 2 Chinstrap                          2.64                       0.169
## 3 Gentoo                             3.20                       0.157
```

Gentoos have the greatest average bill ratio.

Question 6: (2 pts)

Using `dplyr` functions on `penguins_2007`, remove missing values for `sex`. Pipe a `ggplot` to create a single plot showing the distribution of `body_mass_g` colored by male and female penguins, faceted by species (use the function `facet_wrap()` with the option `nrow =` to give each species its own row). Which species shows the least sexual dimorphism (i.e., the greatest overlap of male/female size distributions)?

```
penguins_2007 %>%
  select(body_mass_g, sex, species) %>%    # select relevant info
  filter(!is.na(sex)) %>%    # exclude obs where sex is NA
  ggplot(aes(body_mass_g, sex, fill = sex)) +    #color bars according to sex
  geom_bar(stat="identity") +
  facet_wrap(~species, nrow = 3)
```

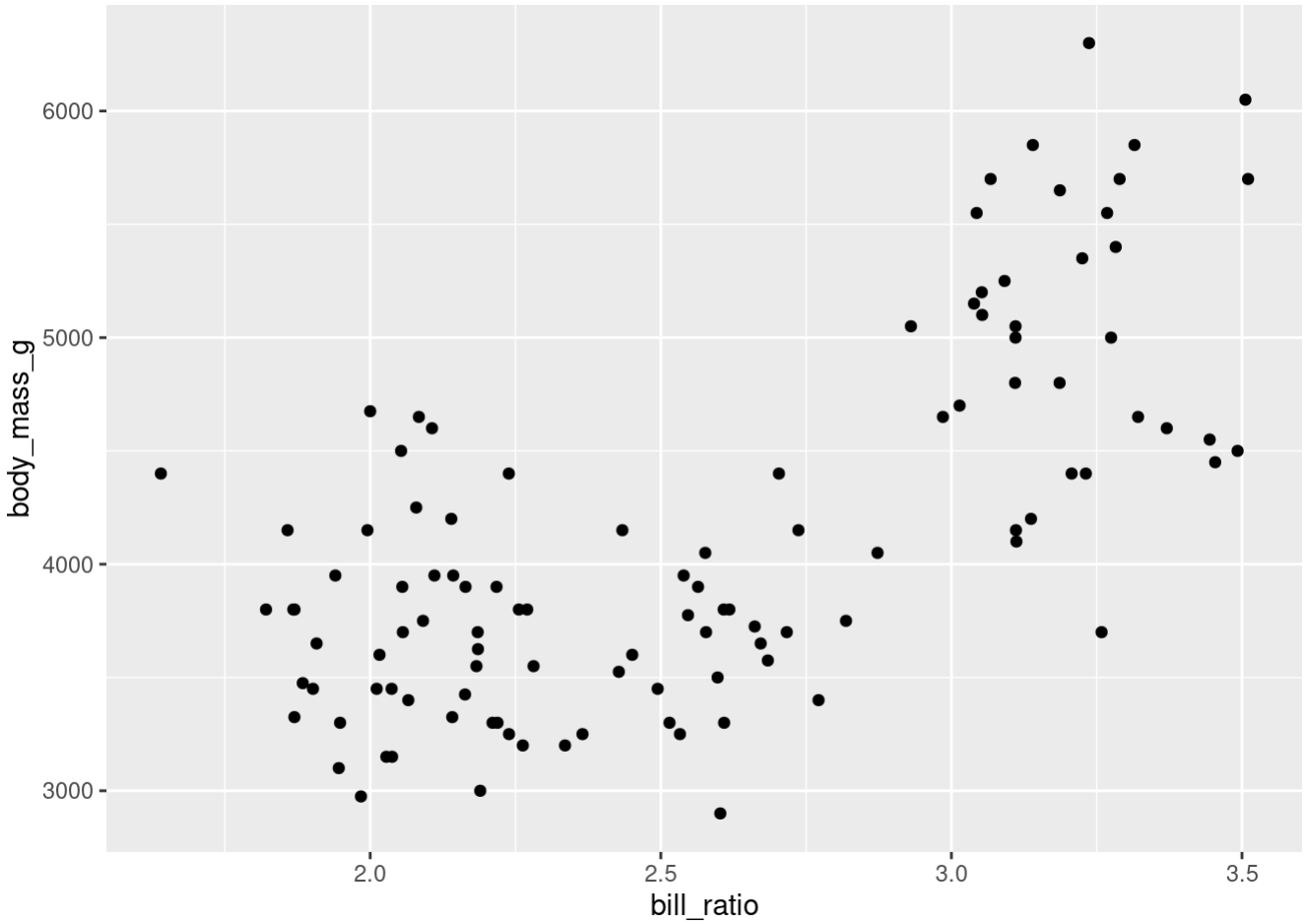


Chinstrap penguins exhibit the least sexual dimorphism; their average body mass is very similar.

Question 7: (2 pts)

Pipe a `ggplot` to `penguins_2007` to create a scatterplot of `body_mass_g` (y-axis) against `bill_ratio` (x-axis). Does it look like there is a relationship between the bill ratio and the body mass? *Note: you might see a Warning message.* What does this message refer to?*

```
penguins_2007 %>%
  mutate(bill_ratio = bill_length_mm / bill_depth_mm, na.rm=T) %>%    # create bill ratio obs
  select(body_mass_g, bill_ratio) %>%    # select relevant obs
  filter(!is.na(bill_ratio) & !is.na(body_mass_g)) %>%    # exclude missing values for either va
riable
  ggplot(aes(bill_ratio, body_mass_g)) +
  geom_point()
```

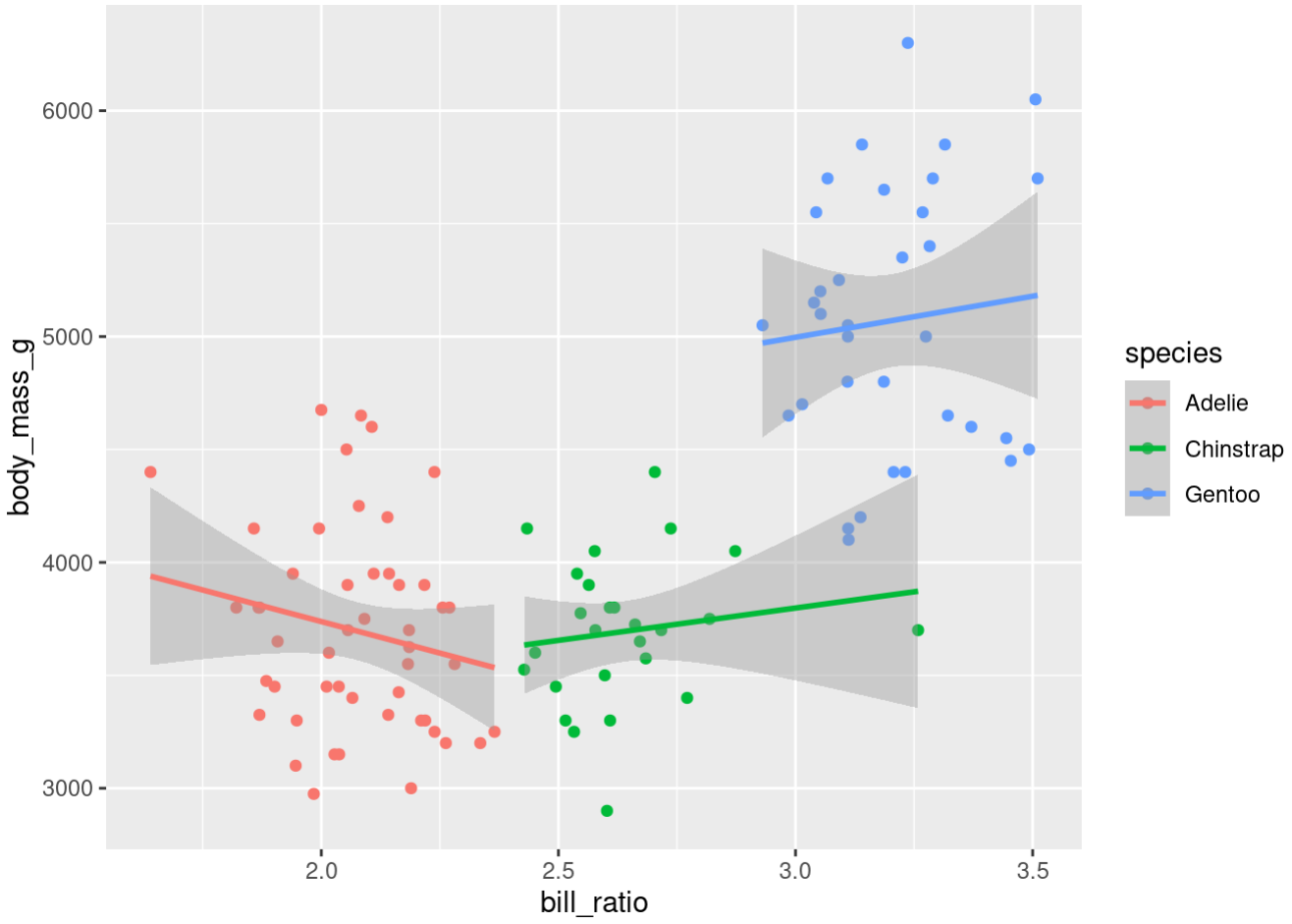


There appears to be a positive linear relationship between body mass and bill ratio. I didn't get a warning message, but I imagine it would be due to missing values in either variable.

Question 8: (2 pts)

What if we separate each species? Duplicate the plot from the previous question and add a regression trend line with `geom_smooth(method = "lm")`. Color the points and the regression lines by species. Does the relationship between the bill ratio and the body mass changes within each species?

```
penguins_2007 %>%
  mutate(bill_ratio = bill_length_mm / bill_depth_mm, na.rm=T) %>%    # create bill ratio obs
  select(body_mass_g, bill_ratio, species) %>%    # select relevant obs
  filter(!is.na(bill_ratio) & !is.na(body_mass_g)) %>%    # exclude missing values for either va
riable
  ggplot(aes(bill_ratio, body_mass_g, color=species)) +
  geom_point() +
  geom_smooth(method = "lm")
```



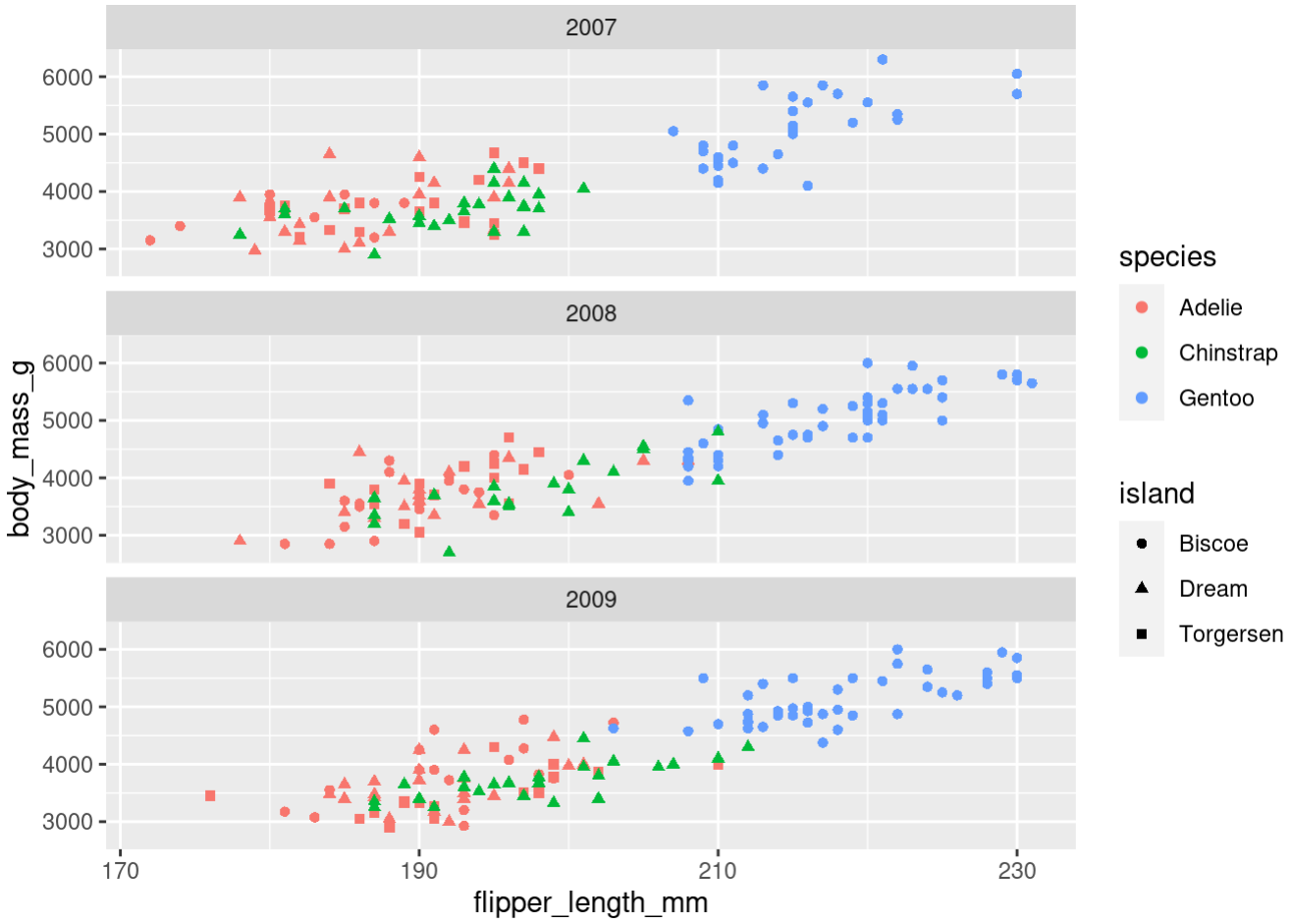
The relationship changes! It appears that Adelie bill ratio increases as body mass decreases.

Question 9: (2 pts)

Finally, let's make a plot using the original `penguins` dataset (not just the 2007 data). Forewarning: This will be very busy plot!

Map `body_mass_g` to the y-axis, `flipper_length_mm` to the x-axis, `species` to color, and `island` to shape. Using `facet_wrap()`, facet the plots by `year`. Find a way to clean up the x-axis labels (e.g., reduce the amount of tick marks) using `scale_x_continuous()`. Does there appear to be a relationship between body mass and flipper length overall? Is there a relationship within each species? What happens to the distribution of flipper lengths for species over time?

```
penguins %>%
  select(body_mass_g, flipper_length_mm, species, island, year) %>%   # select data
  filter(!is.na(body_mass_g) & !is.na(flipper_length_mm)) %>%    # remove missing values
  filter(!is.na(year)) %>%    # more filtering woohoo
  ggplot(aes(flipper_length_mm, body_mass_g, color=species, shape=island)) +
  geom_point() +
  scale_x_continuous(breaks = c(170, 190, 210, 230)) +    # set x-axis breaks
  facet_wrap(~year, nrow = 3)    # distinct year graphs
```



There appears to be a positive linear relationship between body mass and flipper length throughout the dataset; flipper length increases as body mass increases. This holds for all species and across time. The flipper length distribution shifted to the right across time; the penguins' flipper length grew with age.

Formatting: (2 pts)

Comment your code, write full sentences, and knit your file!

```
##                                             sysname
##                                             "Linux"
##                                             release
##                                   "5.15.0-58-generic"
##                                             version
## "#64~20.04.1-Ubuntu SMP Fri Jan 6 16:42:31 UTC 2023"
##                                            nodename
##                           "educcomp01.ccbb.utexas.edu"
##                                             machine
##                                            "x86_64"
##                                               login
##                                           "unknown"
##                                                user
##                                           "jlh7459"
##                                       effective_user
##                                           "jlh7459"
```