HW 4

Enter your name and EID here: Jongho Yoo (jy23294)

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.

```
# filter where year is 2007 and save as new object
penguins_2007 <- penguins %>%
  filter(year == 2007)

# find number of rows in each dataset
nrow(penguins)
```

```
## [1] 344
```

```
nrow(penguins_2007)
```

```
## [1] 110
```

The penguins dataset has 344 rows, while the penguins_2007 dataset has 110 rows.

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?

```
# first group by species and island, then find the counts of observations
penguins_2007 %>%
  group_by(species, island) %>%
  summarize(count = n())
```

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

The species-island observations are: Adelie-Biscoe = 10, Adelie-Dream = 20, Adelie-Torgersen = 20, Chinstrap-Dream = 26, Gentoo-Biscoe = 34. The species Adelie appears 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.

```
# Use the mutate function to create the new bill_ratio variable
penguins_2007 <- penguins_2007 %>%
  mutate(bill_ratio = bill_length_mm/bill_depth_mm)
```

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?

```
# filter for bill_ratio above 3.5 to see if any observations exist
penguins_2007 %>%
  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 species has 2 observations with bill ratio above 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?

```
# first group by species then arrange from smallest to largest
penguins_2007 %>%
  group_by(species) %>%
  arrange(bill_ratio) %>%
  # return the first 3 observations, then select the wanted columns
  slice(1:3) %>%
  select(species, sex, bill_ratio)
```

```
## # A tibble: 9 x 3
## # Groups:   species [3]
##   species   sex    bill_ratio
##   <fct>     <fct>       <dbl>
## 1 Adelie    male         1.64
## 2 Adelie    male         1.82
## 3 Adelie    male         1.86
## 4 Chinstrap female       2.43
## 5 Chinstrap female       2.43
## 6 Chinstrap female       2.45
## 7 Gentoo    male         2.93
## 8 Gentoo    female       2.99
## 9 Gentoo    female       3.01
```

No. While the Adelie and Chinstrap species have the same sex as the 3 smallest bill ratios, the Gentoo species has male and females in the 3 smallest bill ratios.

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`?

```
# first group by species, then use summarize function to find mean and sd of bill ratio
penguins_2007 %>%
  group_by(species) %>%
  summarize(mean_bill_ratio = mean(bill_ratio, na.rm = T),
            sd_bill_ratio = sd(bill_ratio, na.rm = T))
```

```
## # A tibble: 3 x 3
##   species   mean_bill_ratio sd_bill_ratio
##   <fct>               <dbl>         <dbl>
## 1 Adelie               2.07         0.152
## 2 Chinstrap            2.64         0.169
## 3 Gentoo               3.20         0.157
```

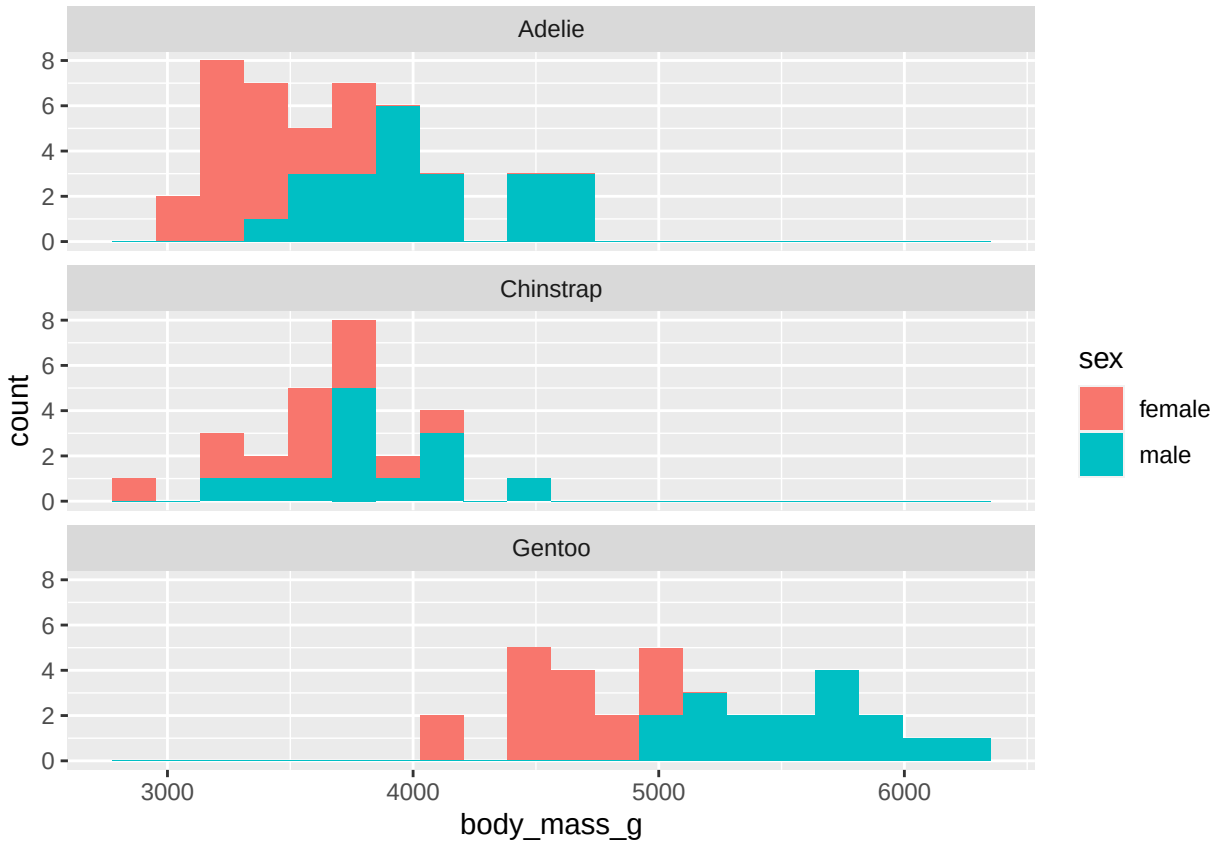
The Gentoo species has the greatest mean bill ratio of 3.20mm

---

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)?

```
# first filter out NA values for sex, then plot the distribution of body mass between
# males and females per species
penguins_2007 %>%
  filter(!is.na(sex)) %>%
  ggplot(aes(body_mass_g, fill = sex)) +
  geom_histogram(bins = 20) +
  facet_wrap(~species, nrow = 3)
```

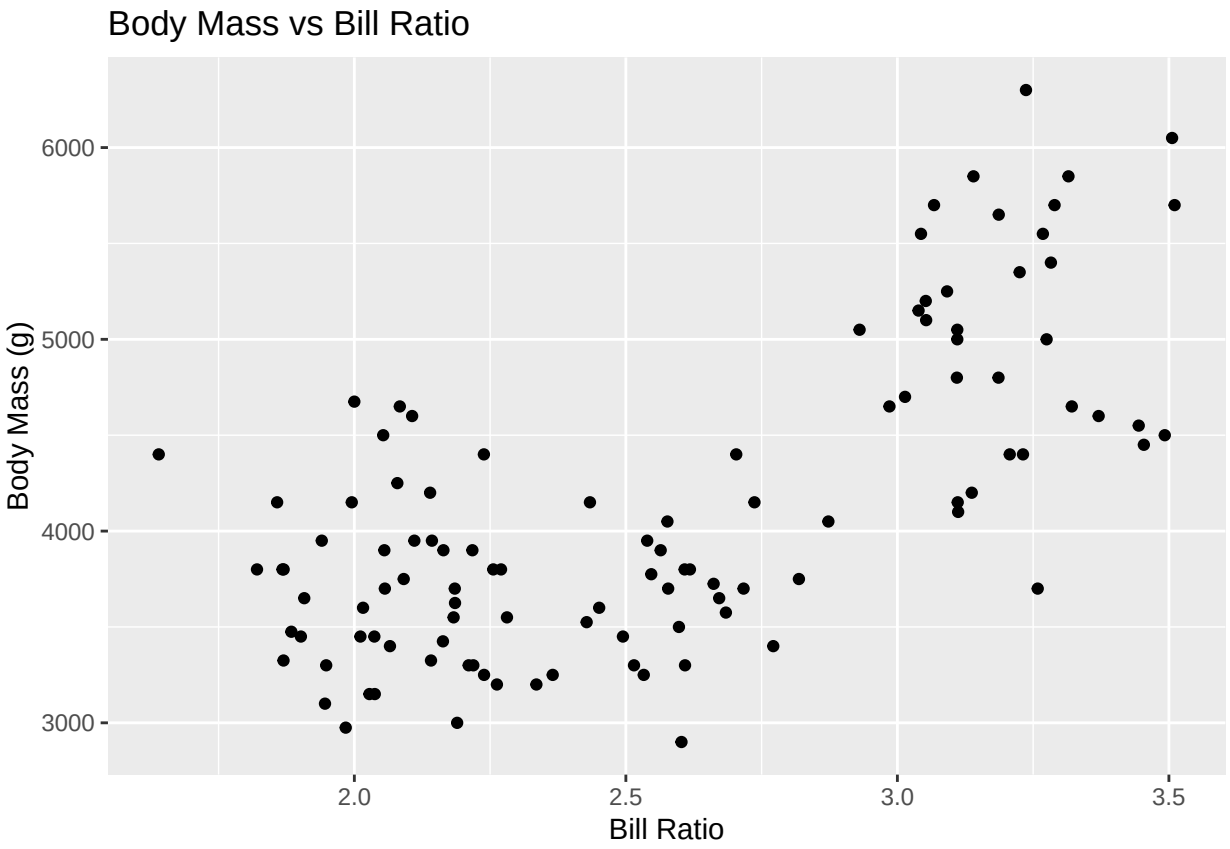


The Chinstrap species shows the least sexual dimorphism.

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?*

```
# create a scatterplot of body mass vs bill ratio
penguins_2007 %>%
  ggplot(aes(x = bill_ratio, y = body_mass_g)) +
  geom_point() +
  labs(title = "Body Mass vs Bill Ratio",
       x = "Bill Ratio",
       y = "Body Mass (g)")
```

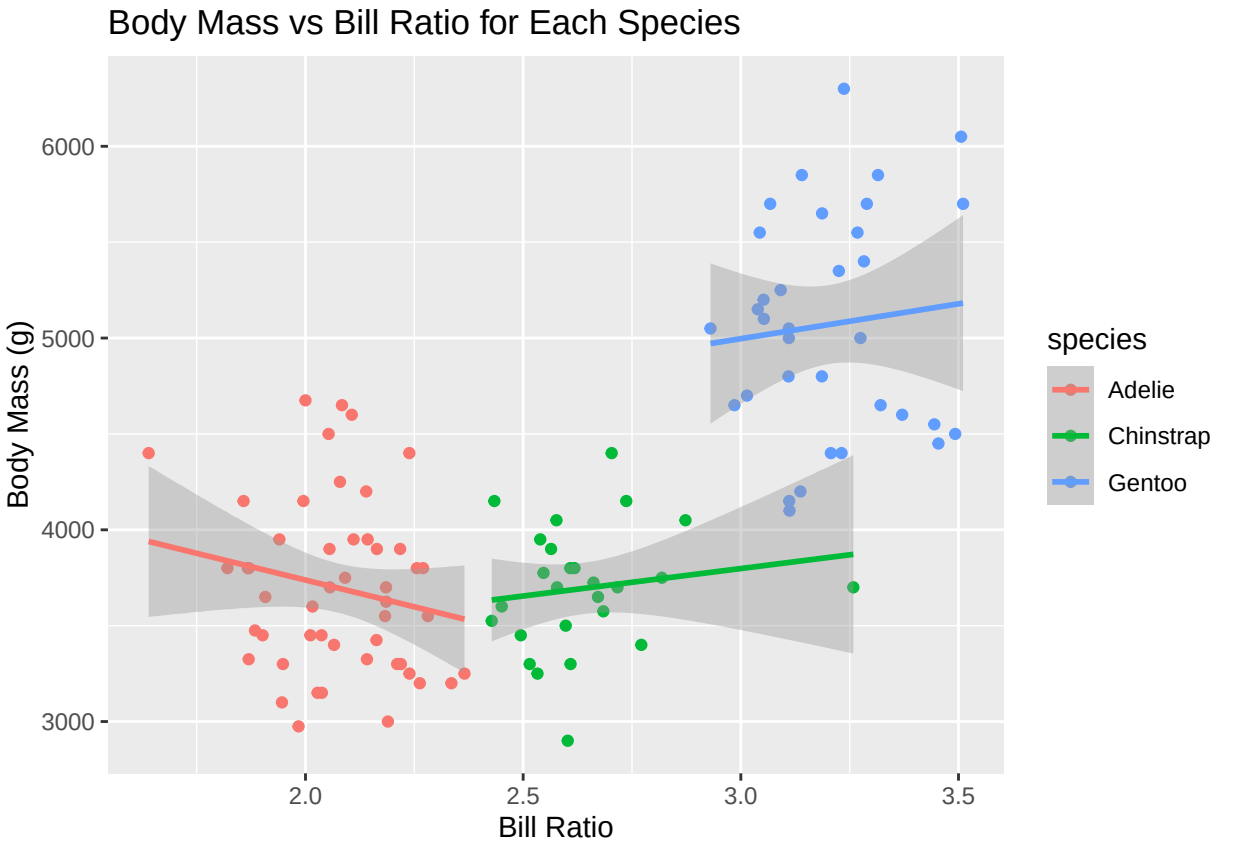


Yes, there seems to be a positive correlation between body mass and bill ratio. The error message refers to the NA value.

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?

```r
# Group the plot from Q7 by species, then fit a linear regression line for each species
penguins_2007 %>%
  group_by(species) %>%
  ggplot(aes(x = bill_ratio, y = body_mass_g, color = species)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Body Mass vs Bill Ratio for Each Species",
       x = "Bill Ratio",
       y = "Body Mass (g)")
```



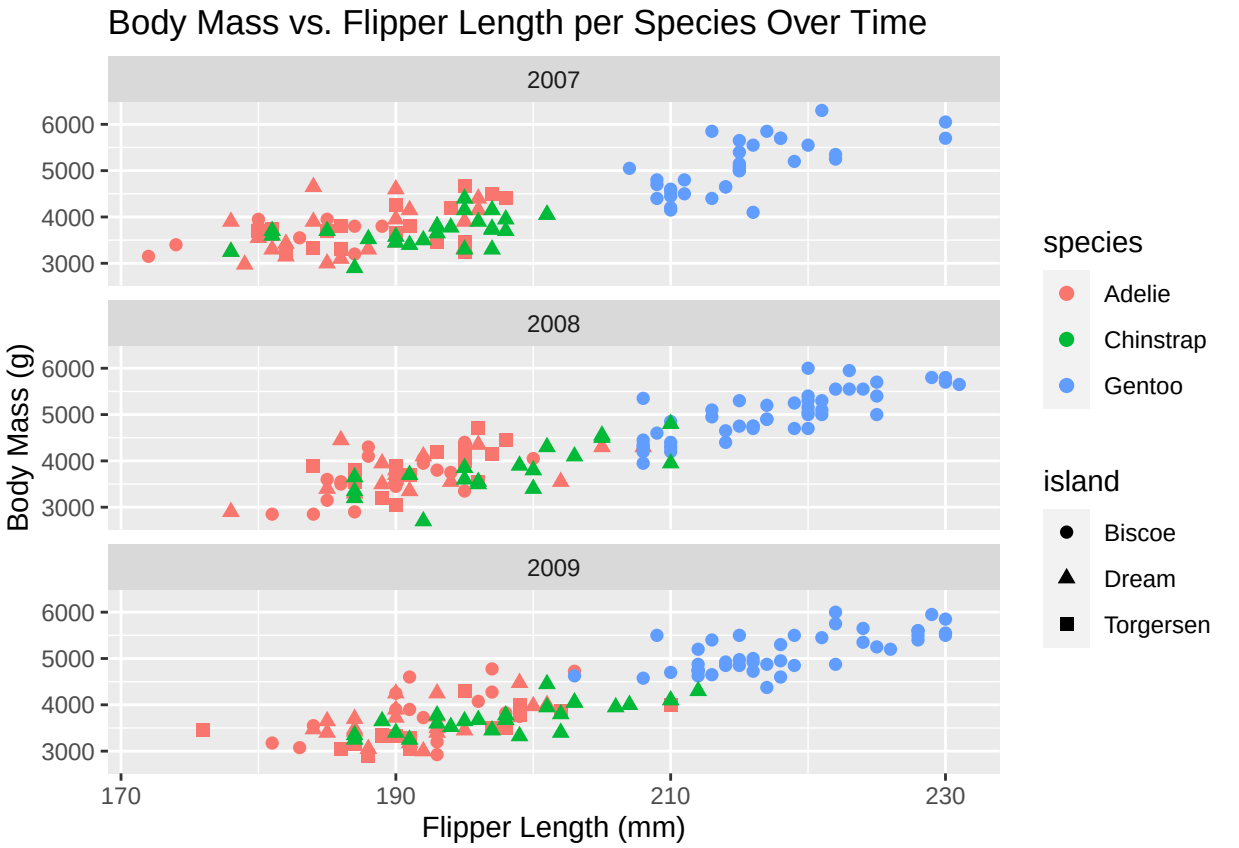
Yes, whereas the Chinstrap and Gentoo species have a slightly positive correlation, the Adelie species has a slightly negative correlation.

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?

```r
# Plot flipper length vs. body mass per species and island
penguins %>%
  ggplot(aes(x = flipper_length_mm, y = body_mass_g, color = species, shape = island)) +
  geom_point(size = 2) +
  # facet wrap the plot to separate by year
  facet_wrap(~ year, nrow = 3) +
  scale_x_continuous(breaks = seq(170, 240, 20)) +
  labs(title = "Body Mass vs. Flipper Length per Species Over Time",
       x= "Flipper Length (mm)",
       y = "Body Mass (g)")
```



The distribution of flipper length seems to slightly shift right (increase) over time. We can see how in 2007, the Adelie and Chinstrap species flipper length distribution was roughly split evenly between the 190mm x-line. However, in 2009, most of the distribution is past the 190mm x-line, indicating a shift (increase) in flipper length distribution.

Formatting: (2 pts)

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

```
##                                                                                                                         sysn
##                                                                                                                        "Darw
##                                                                                                                         rele
##                                                                                                                        "21.3
##                                                                                                                         vers
## "Darwin Kernel Version 21.3.0: Wed Jan  5 21:37:58 PST 2022; root:xnu-8019.80.24~20/RELEASE_ARM64_T8
##                                                                                                                        noden
##                                                                               "wireless-10-146-190-14.public.utexas.e
##                                                                                                                         mach
##                                                                                                                         "arm
##                                                                                                                           lo
##                                                                                                                          "ro
##                                                                                                                            u
##                                                                                                                     "steveny
##                                                                                                                   effective_u
##                                                                                                                     "steveny
```