HW 4

Enter your name and EID here: Erik Mercado, emm4376

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 for data only from 2007
penguins_2007 <- penguins |>
  filter(year==2007)
```

Penguins has 344 observations and penguins_2007 has 110 observations.

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?

```
# group by species and then by island. Count the totals per group combo.
penguins_2007 |>
  group_by(species,island) |>
  summarise(n=n())
```

```
## # A tibble: 5 x 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
```

There are 10 Adelie living on Biscoe. There are 20 Adelie living on Dream. There are 20 Adelie living on Torgersen. There are 26 Chinstrap living on Dream. There are 34 Gentoo living on Biscoe.

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.

```
# create new variable called "bill_ratio"
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 to only get penguins with a bill ration > 3.5
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
```

There are two penguins with a bill ratio greater than 3.5. They are both from the Gentoo species.

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?

```r
# group by species, them arrange by bill ratio, then print out only bottom 3, then select desired colum
penguins_2007 |>
  group_by(species)|>
  arrange(bill_ratio) |>
  top_n(-3, bill_ratio) |>
  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, the same sex does not have the smallest bill ratio across species.

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`?

```r
# get the mean and sd per species (excluding NA values)
penguins_2007 |>
  group_by(species) |>
  summarize(mean_bill_ratio = mean(bill_ratio, na.rm = T), sd_bill_ratio = sd(bill_ratio, na.rm = T))
```

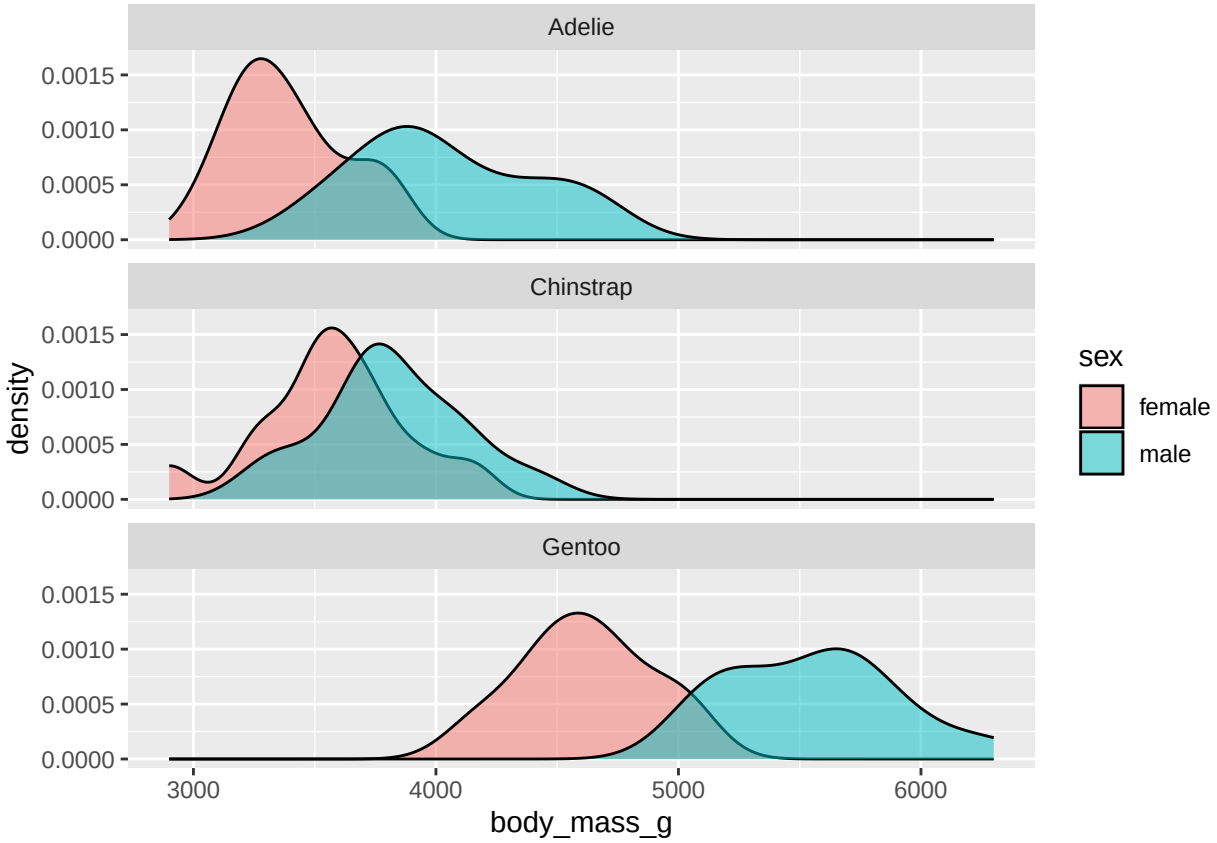
```
## # A tibble: 3 x 3
##   species   mean_bill_ratio sd_bill_ratio
##   <fct>               <dbl>         <dbl>
## 1 Adelie               2.07         0.152
## 2 Chinstrap            2.64         0.169
## 3 Gentoo               3.20         0.157
```

For the bill ratio of Adelie, the mean is 2.074500 and the standard deviation is 0.1515183. For the bill ratio of Chinstrap, the mean is 2.638122 and the standard deviation is 0.1694886. For the bill ratio of Gentoo, the mean is 3.203209 and the standard deviation is 0.1565563.

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)?

```
# filter out na's in sex and then create a distribution divided by sex
penguins_2007 |>
  filter(!is.na(sex)) |>
  ggplot(aes(x=body_mass_g, fill=sex)) +
  geom_density(alpha=.5) +
  facet_wrap(vars(species),nrow=3)
```

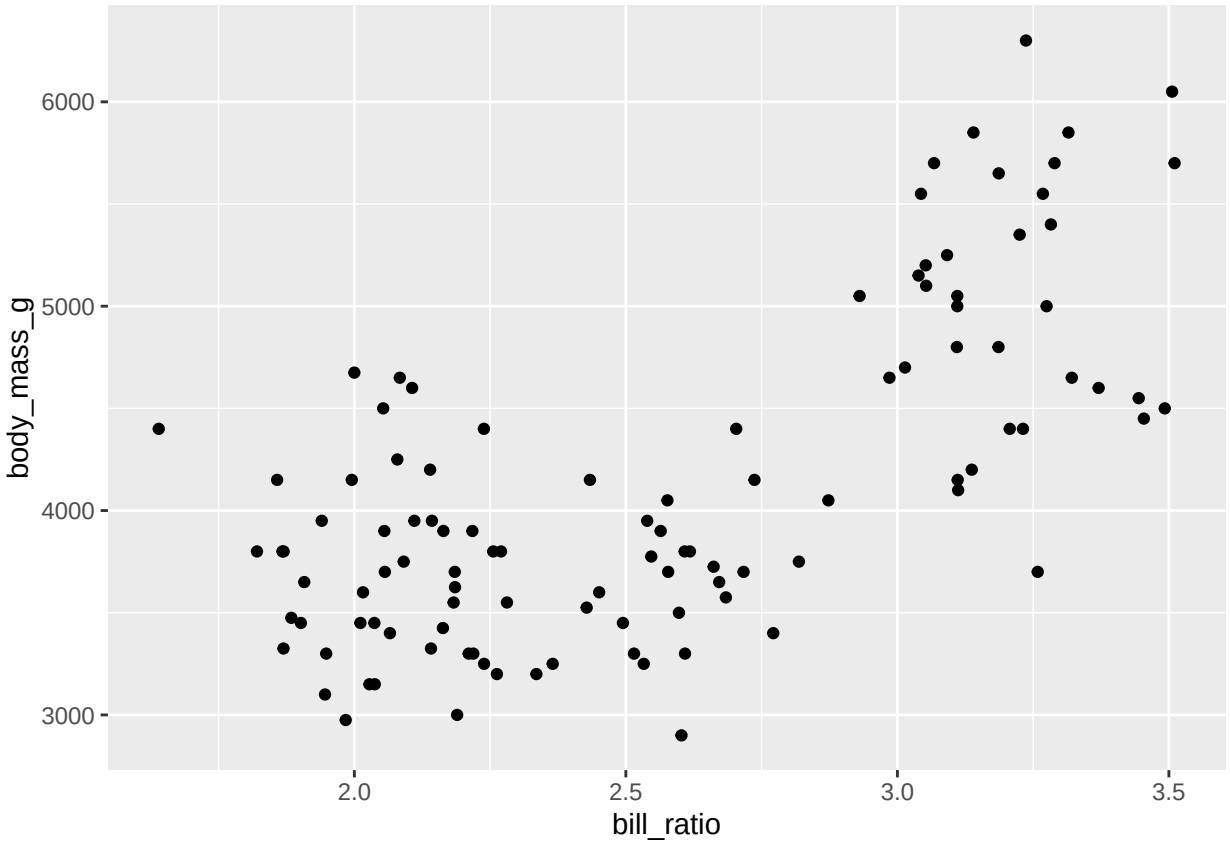


It seems that the Gentoo species has the least sexual dimorphism.

Question 7: (2 pts)

Pipe`penguins_2007` to `ggplot()` 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?*

```r
# create scatter plot of bill_ratio against body_mass_g
penguins_2007 |>
  ggplot(aes(x=bill_ratio, y=body_mass_g)) +
  geom_point()
```

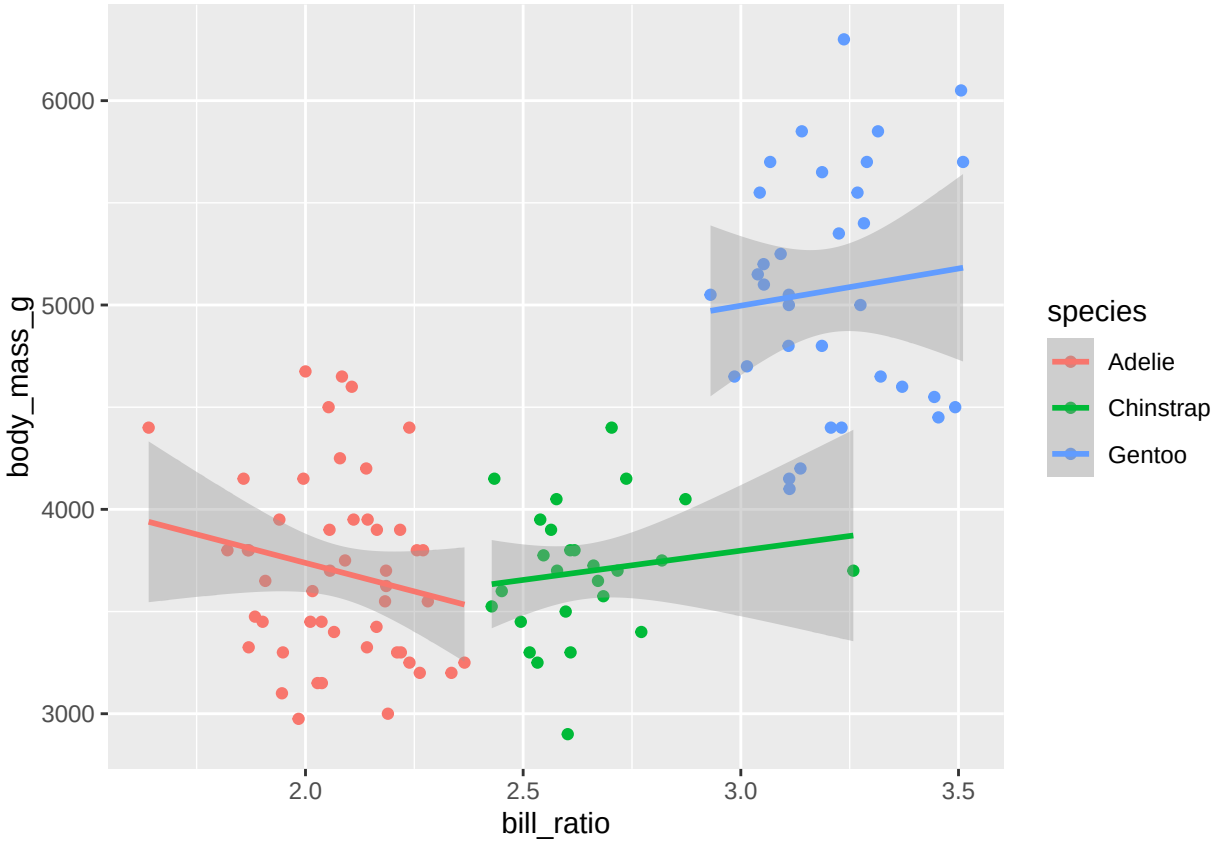


There does seem to be a bit of a positive correlation between body_mass_g and _bill_ratio. The row is saying that geom_point removed 1 row because it was missing data for one or both of the variables used in the graph. Looking through the data, it is talking about observation 4.

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 appear to be the same across the different species?

```r
# breakdown previos scatter plot by species
penguins_2007 |>
  ggplot(aes(x=bill_ratio, y=body_mass_g)) +
  geom_point(aes(color=species)) +
  geom_smooth(method = "lm",aes(color=species))
```



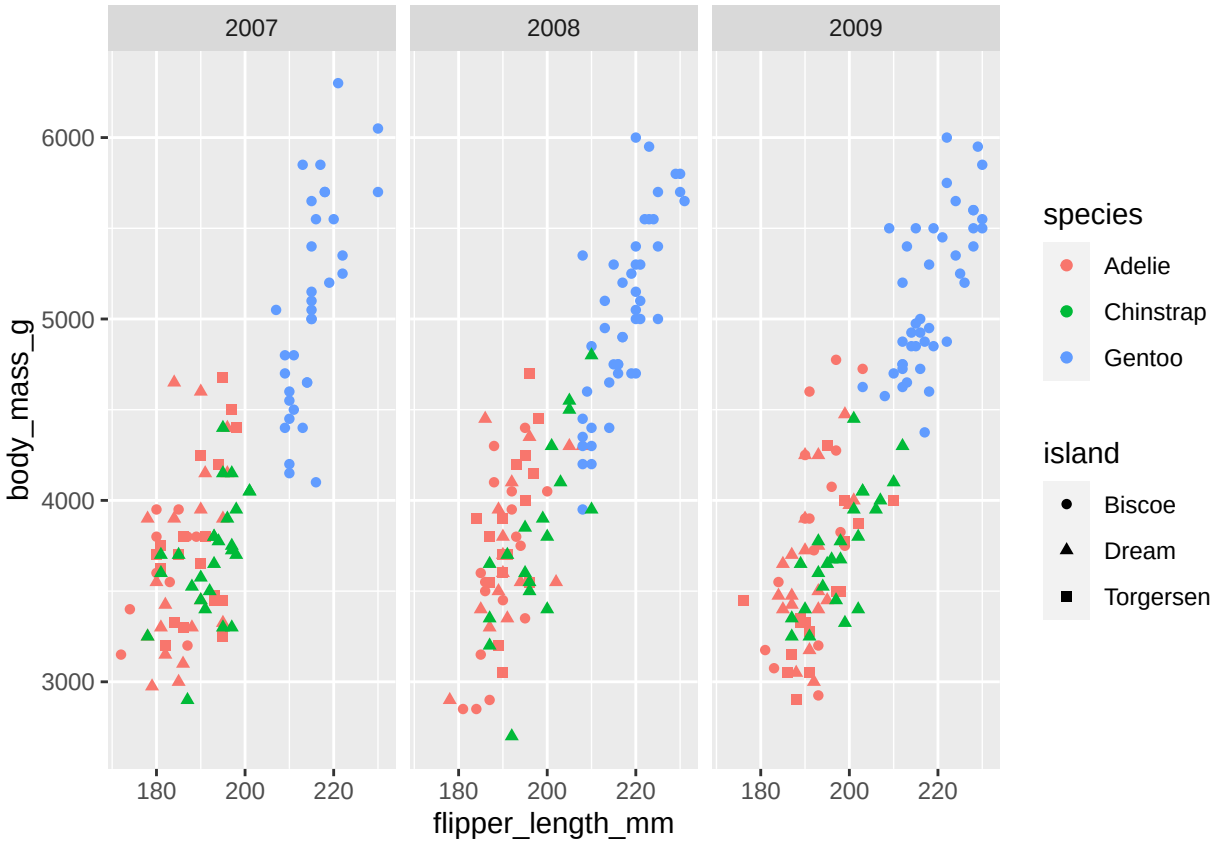
The relationship between bill_ratio and body_mass_g does not appear to be the same across all species. It is positive with the Chinstrap and Gentoo species, but negative with the Adelie species.

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 number 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?

```
# create a scatter plot divided up by species,island, and year
penguins |>
  ggplot(aes(x=flipper_length_mm, y=body_mass_g, color=species, shape = island)) +
  geom_point() +
  facet_wrap(vars(year)) +
  scale_x_continuous(breaks=seq(160,240,by=20))
```



The overall relationship seems to be positive between flipper_length_mm and body_mass_g. This relationship also carries over into the individual species.

Formatting: (2 pts)

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