HW 4

Enter your name and EID here: Harini Shanmugam hs28663

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/observations 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.

```
# storing information from from pengins data set from just 2007 into new object
penguins_2007 <- penguins %>%
  filter(year==2007)

nrow(penguins)
```

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

```
nrow(penguins_2007)
```

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

The original penguins dataset has 344 observations. The new penguins_2007 dataset 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?

```
penguins_2007 %>%
  group_by(species, island) %>% # grouping by species, then by island
  summarize(count=n()) # count per species-island
```

```
## # 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
```

Number of observations for each species-island combination:

Adelie-Torgersen: 10

Adelie-Biscoe: 20

Adelie-Dream: 20

Chinstrap-Dream: 26

Gentoo-Biscoe: 34

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.

```
# create new variable bill_ratio by algebra (division) of existing two variables
# and add it into penguins_2007 dataset
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 that is greater than 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
```

Yes, there are 2 cases in the penguins_2007 dataset for which the bill_ratio exceeds 3.5 and they are of 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?

```
# 3 smallest bill ratios of each species
penguins_2007 %>%
  group_by(species) %>%
  slice_min(bill_ratio, n=3) %>% # 3 smallest for each species
  select(species, sex, bill_ratio) # display only these variables
```

```
## # 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
```

The same sex does not have the smallest bill ratio across the species. The smallest bill ratios for the Adelie species belong only to males. The smallest bill ratios for the Chinstrap species belong only to females. However, the smallest bill ratios for the Gentoo species belong to both males and females.

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`?

```
# mean and standard deviation of each of the species
penguins_2007 %>%
  group_by(species) %>%
  summarize(mean_bill_ratio = mean(bill_ratio, na.rm=T), # calculate mean
            sd_bill_ratio = sd(bill_ratio, na.rm=T)) # calculate sd
```

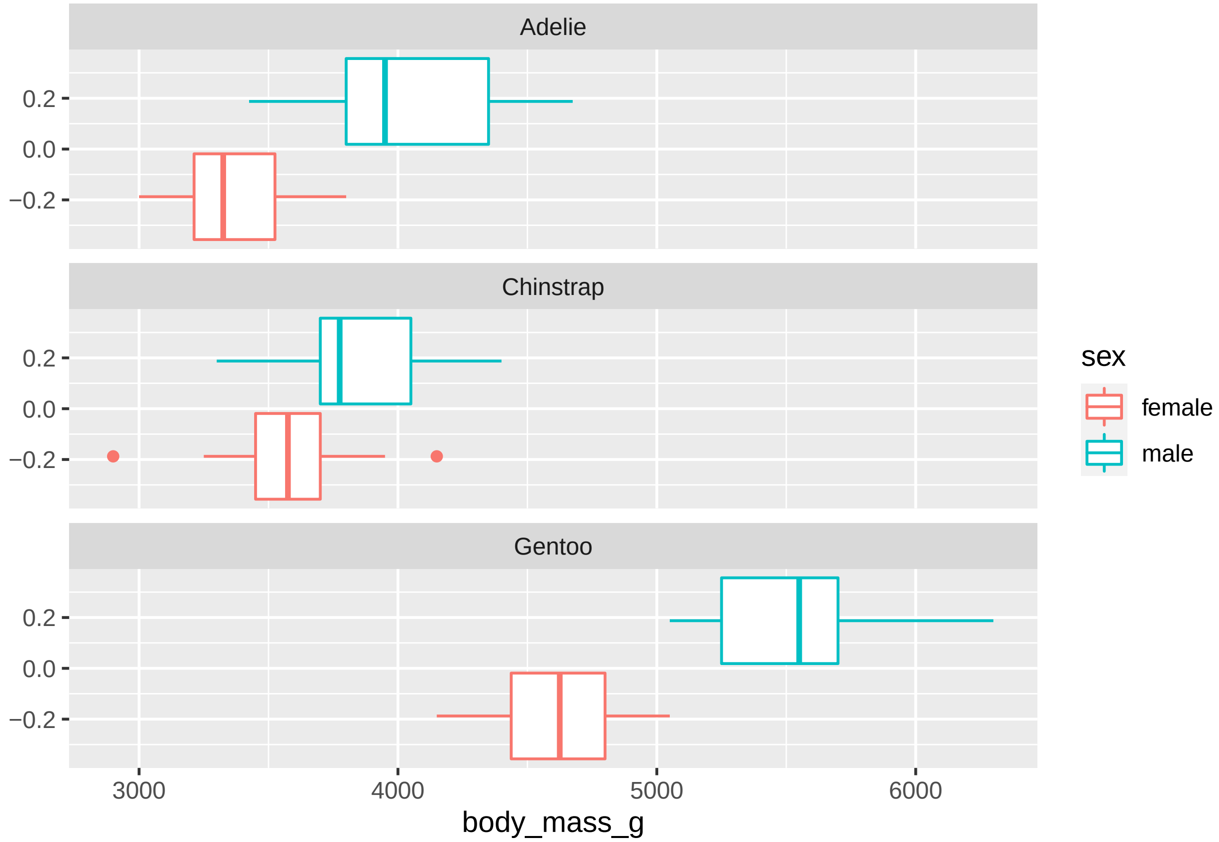
```
## # A tibble: 3 x 3
##   species   mean_bill_ratio sd_bill_ratio
##   <fct>               <dbl>         <dbl>
## 1 Adelie               2.07         0.152
## 2 Chinstrap            2.64         0.169
## 3 Gentoo               3.20         0.157
```

The Gentoo species has the greatest mean bill_ratio.

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 %>%
  filter(!is.na(sex)) %>% # remove NA values in sex variable
  ggplot(aes(x = body_mass_g, color = sex)) +
  geom_boxplot() +
  facet_wrap(~ species, nrow=3) # separate graphs by species type and stack
```

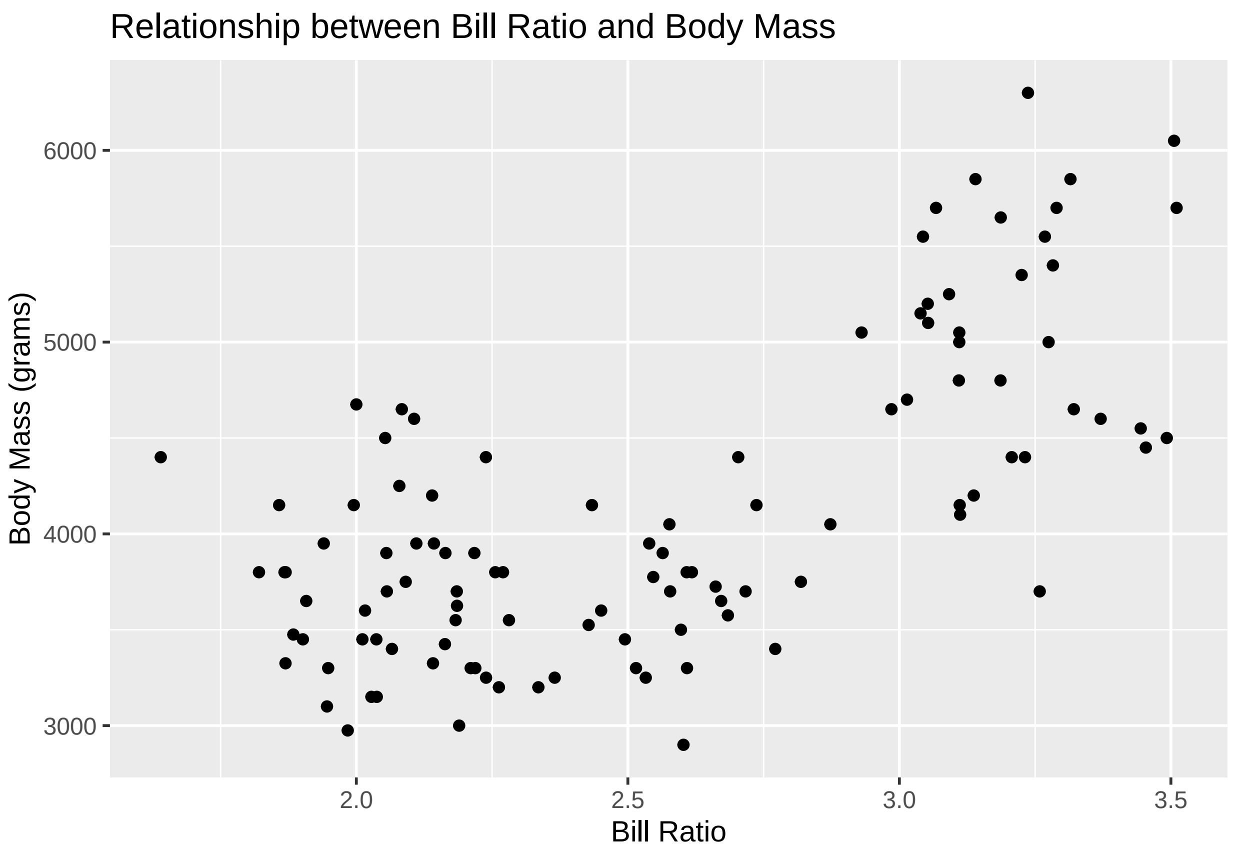


The Chinstrap species shows the least sexual dimorphism.

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?*

```
# scatter plot of bill ratio vs body mass
penguins_2007 %>%
  ggplot(aes(x=bill_ratio, y=body_mass_g)) +
  geom_point() +
  labs(title="Relationship between Bill Ratio and Body Mass",
       x="Bill Ratio",
       y="Body Mass (grams)")
```

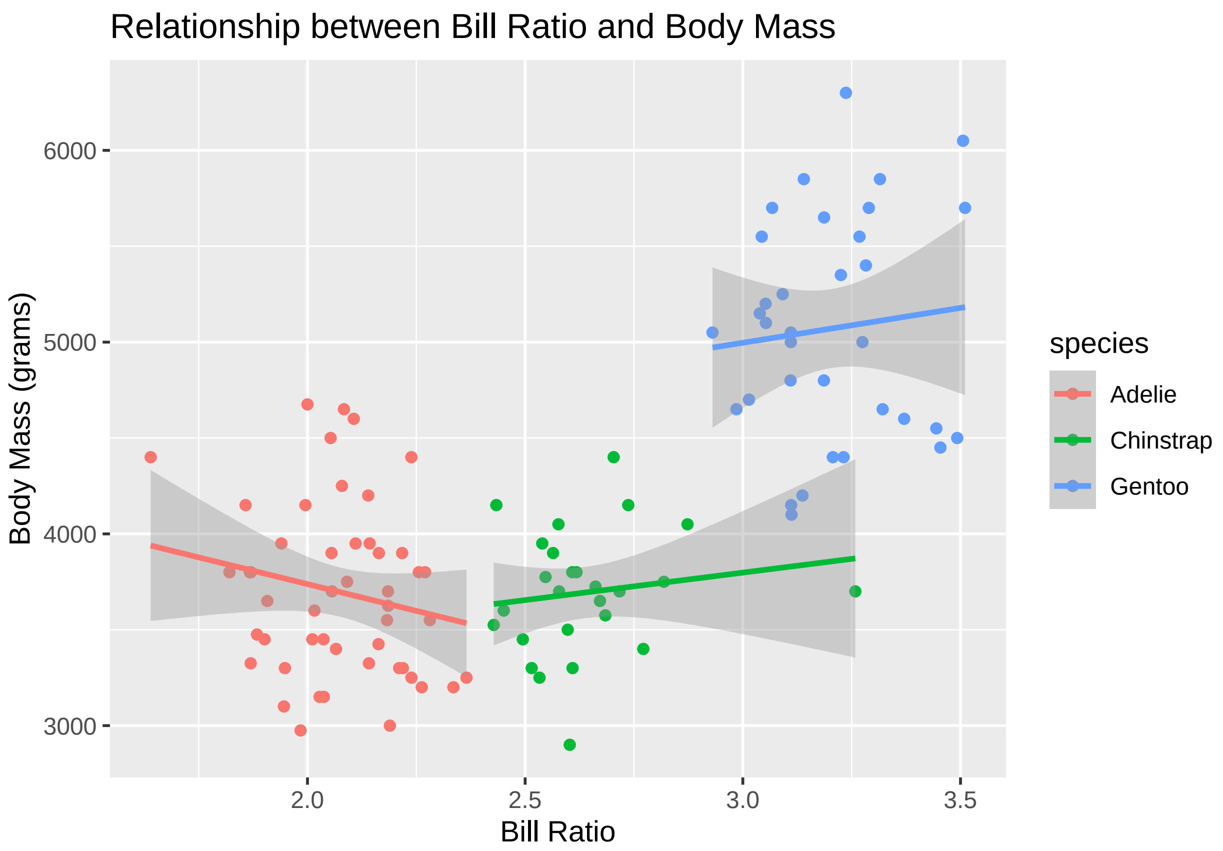


Yes, there looks like there is a positive linear relationship between bill ratio and body mass. The warning message refers to the one row with NA values.

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
# make scatter plot of bill ratio vs body mass
penguins_2007 %>%
  ggplot(aes(x=bill_ratio, y=body_mass_g, color=species)) +
  geom_point() +
  geom_smooth(method = "lm") + # regression line
  labs(title="Relationship between Bill Ratio and Body Mass",
       x="Bill Ratio",
       y="Body Mass (grams)")
```



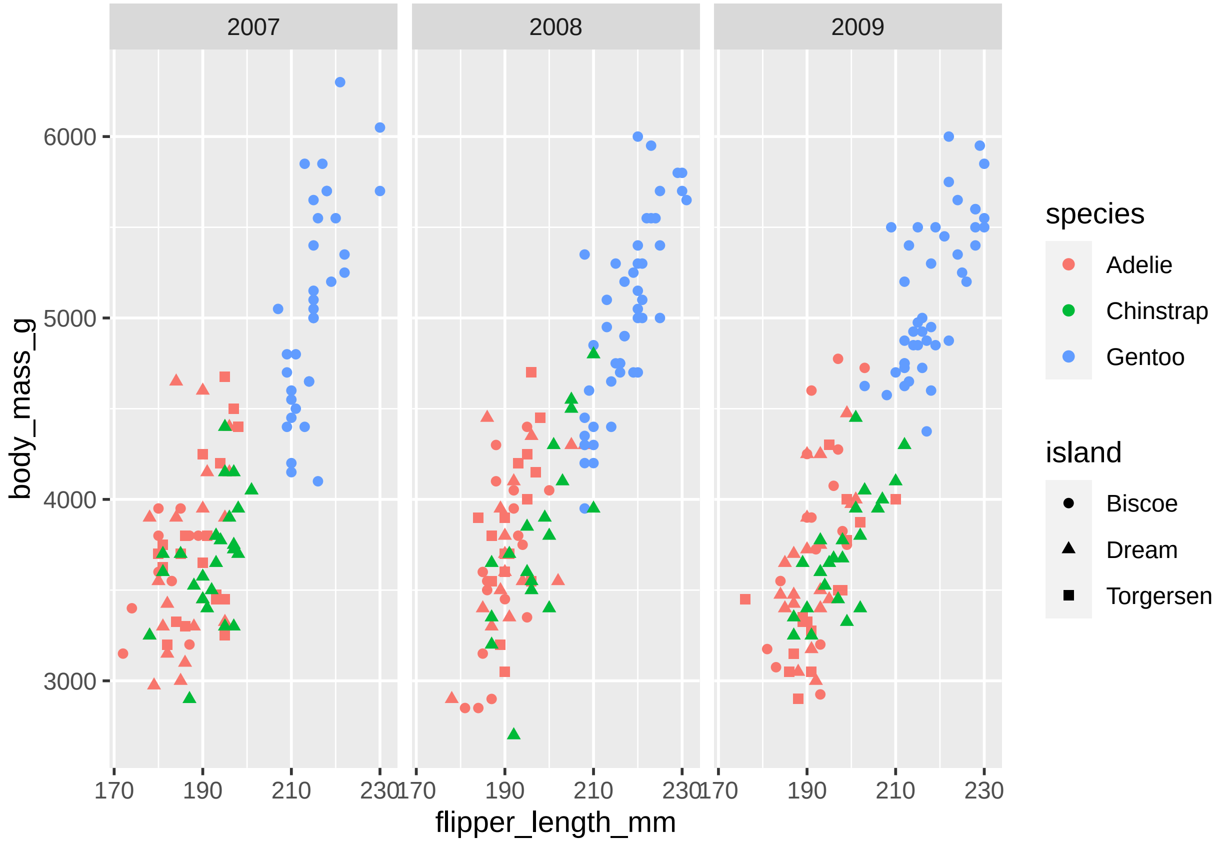
Unlike the Chinstrap and Gentoo species which have a positive linear correlation between bill ratio and body mass, the Adelie species has a negative linear relationship, meaning that the greater the bill ratio, the smaller the body mass.

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
# scatter plot of flipper length vs body mass
penguins %>%
  ggplot(aes(x=flipper_length_mm, y=body_mass_g, color=species, shape=island)) +
  geom_point() +
  facet_wrap(~ year) + # separate by year
  # x-axis ticks to go from 170 to 230 by margins of 20
  scale_x_continuous(breaks = seq(170,230,20))
```



Yes there seems to be a positive linear relationship between body mass and flipper length: the longer the flipper length, the higher the body mass. Within each species, the Adelie and Chinstap species had similar flipper lengths and body mass. The Gentoo species had the highest flipper lengths and highest body mass. Over time, the spread (of the body mass) decreased.

Formatting: (2 pts)

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

```
##                                                                                                 sysna
##                                                                                                "Darwi
##                                                                                                 relea
##                                                                                                "21.6
##                                                                                                 vers
## "Darwin Kernel Version 21.6.0: Wed Aug 10 14:28:35 PDT 2022; root:xnu-8020.141.5~2/RELEASE_ARM64_T81
##                                                                                                nodena
##                                                                              "Harinis-Air.attlocal.n
##                                                                                                 machi
##                                                                                                 "arm6
##                                                                                                        log
##                                                                                                       "roo
##                                                                                                         us
##                                                                                           "harinishanmuga
##                                                                                               effective_us
##                                                                                           "harinishanmuga
```