HW 5

SDS322E

September 28, 2022

Ethan Chang - ehc586

Please submit as a PDF or HTML file on Canvas before the due date.

For all questions, include the R commands/functions that you used to find your answer. Answers without supporting code will not receive credit.

> Review of how to submit this assignment
>
> All homework assignments will be completed using R Markdown. These `.Rmd` files consist of >text/syntax (formatted using Markdown) alongside embedded R code. When you have completed the assignment (by adding R code inside codeblocks and supporting text outside of the codeblocks), create your document as follows (assuming you are using the edupod server and submitting HTML):

- Click the arrow next to the “Knit” button (above)
- Choose “Knit to HTML”
- Go to Files pane and put checkmark next to the correct HTML file
- Click on the blue gear icon (“More”) and click Export
- Download the file and then upload to Canvas
- To submit a PDF, open your HTML file and print it to a pdf, then upload the pdf as your submission.

Question 1 (1 pts)

In this homework you will practice your `dplyr` chops on the `penguins` dataset, which is inside the `palmerpenguins` package (you will need to run `install.packages("palmerpenguins")` if the package is not installed already), we we can grab this as well:

```
library(tidyverse)
library(palmerpenguins)
```

Read the documentation by running `?penguins` to familiarize yourself with the columns.

Now, use `filter()` to pick all the rows/observations in the `penguins` dataset from the year 2007 and store them in a new object called `penguins_2007`. Then compare the number of rows in the original `penguins` dataset with your new dataset in words.

```
data(penguins)
penguins_2007 <- filter(penguins, year == 2007)
str(penguins)
```

```
## tibble [344 × 8] (S3: tbl_df/tbl/data.frame)
##  $ species          : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ island           : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ bill_length_mm   : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
##  $ bill_depth_mm    : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
##  $ flipper_length_mm: int [1:344] 181 186 195 NA 193 190 181 195 193 190 ...
##  $ body_mass_g      : int [1:344] 3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 ...
##  $ sex              : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2 NA NA ...
##  $ year             : int [1:344] 2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
```

```
str(penguins_2007)
```

```
## tibble [110 × 8] (S3: tbl_df/tbl/data.frame)
##  $ species          : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ island           : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3 ...
##  $ bill_length_mm   : num [1:110] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
##  $ bill_depth_mm    : num [1:110] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
##  $ flipper_length_mm: int [1:110] 181 186 195 NA 193 190 181 195 193 190 ...
##  $ body_mass_g      : int [1:110] 3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 ...
##  $ sex              : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2 NA NA ...
##  $ year             : int [1:110] 2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
```

Answer: It can be seen that there are 344 rows in the original `penguins` dataset compared to the 110 rows in the new `penguins_2007` dataset. There are more rows in the original dataset, which makes sense as our new dataset is extracted from this original, so it is natural that this new one will have less rows.

Question 2 (0.5 pts)

Return all the rows in `penguins_2007` where `bill_length_mm` is between 45 and 55 (doesn't matter if we include 45 or 55 specifically, since no observations have those exact values).

```
penguins_2007 %>%
    filter(bill_length_mm < 55, bill_length_mm > 45)
```

```
## # A tibble: 51 × 8
##    species island    bill_length_mm bill_depth_mm flipper_…¹ body_…² sex    year
##    <fct>   <fct>              <dbl>         <dbl>      <int>   <int> <fct> <int>
##  1 Adelie  Torgersen           46            21.5        194    4200 male   2007
##  2 Gentoo  Biscoe              46.1          13.2        211    4500 fema…  2007
##  3 Gentoo  Biscoe              50            16.3        230    5700 male   2007
##  4 Gentoo  Biscoe              48.7          14.1        210    4450 fema…  2007
##  5 Gentoo  Biscoe              50            15.2        218    5700 male   2007
##  6 Gentoo  Biscoe              47.6          14.5        215    5400 male   2007
##  7 Gentoo  Biscoe              46.5          13.5        210    4550 fema…  2007
##  8 Gentoo  Biscoe              45.4          14.6        211    4800 fema…  2007
##  9 Gentoo  Biscoe              46.7          15.3        219    5200 male   2007
## 10 Gentoo  Biscoe              46.8          15.4        215    5150 male   2007
## # … with 41 more rows, and abbreviated variable names ¹flipper_length_mm,
## #   ²body_mass_g
## # ℹ Use `print(n = ...)` to see more rows
```

Question 3 (0.5 pts)

Are there any cases in `penguins_2007` for which the ratio of `bill_length_mm` to `bill_depth_mm` exceeds 3.5? For now, use only `filter()` to find out. If so, for which species of penguins is this true?

```
penguins_2007 %>%
    filter(bill_length_mm/bill_depth_mm > 3.5)
```

```
## # A tibble: 2 × 8
##   species island bill_length_mm bill_depth_mm flipper_leng…¹ body_…² sex    year
##   <fct>   <fct>           <dbl>         <dbl>          <int>   <int> <fct> <int>
## 1 Gentoo  Biscoe           50.2          14.3            218    5700 male   2007
## 2 Gentoo  Biscoe           59.6          17              230    6050 male   2007
## # … with abbreviated variable names ¹flipper_length_mm, ²body_mass_g
```

Answer: There are 2 cases in `penguins_2007` for which the ratio of `bill_length_mm` to `bill_depth_mm` exceeds 3.5. This is only true for the Gentoo species of penguins.

Question 4 (0.5 pts)

Take your `penguins_2007` dataset and, using `select()`, drop/delete the column `year`. Store the result in the new dataset `penguins_2007_ny`.

```
penguins_2007_ny <- penguins_2007 %>%
    select(-year)
penguins_2007_ny
```

```
## # A tibble: 110 × 7
##    species island    bill_length_mm bill_depth_mm flipper_length…¹ body_…² sex
##    <fct>   <fct>              <dbl>         <dbl>            <int>   <int> <fct>
##  1 Adelie  Torgersen           39.1          18.7              181    3750 male
##  2 Adelie  Torgersen           39.5          17.4              186    3800 fema…
##  3 Adelie  Torgersen           40.3          18                195    3250 fema…
##  4 Adelie  Torgersen           NA            NA                 NA      NA <NA>
##  5 Adelie  Torgersen           36.7          19.3              193    3450 fema…
##  6 Adelie  Torgersen           39.3          20.6              190    3650 male
##  7 Adelie  Torgersen           38.9          17.8              181    3625 fema…
##  8 Adelie  Torgersen           39.2          19.6              195    4675 male
##  9 Adelie  Torgersen           34.1          18.1              193    3475 <NA>
## 10 Adelie  Torgersen           42            20.2              190    4250 <NA>
## # … with 100 more rows, and abbreviated variable names ¹flipper_length_mm,
## #   ²body_mass_g
## # ℹ Use `print(n = ...)` to see more rows
```

Question 5 (0.5 pts)

Using the `mutate()` function, take `penguins_2007_ny` and create a new data column that contains the ratio of `bill_length_mm` to `bill_depth_mm` (call it `bill_ratio`). Write the result to the new dataset `penguins_2007_br`.

```
penguins_2007_br <- penguins_2007_ny %>%
    mutate(bill_ratio = bill_length_mm/bill_depth_mm)
penguins_2007_br
```

```
## # A tibble: 110 × 8
##    species island    bill_length_mm bill_depth_mm flippe…¹ body_…² sex   bill_…³
##    <fct>   <fct>              <dbl>         <dbl>    <int>   <int> <fct>   <dbl>
##  1 Adelie  Torgersen           39.1          18.7      181    3750 male     2.09
##  2 Adelie  Torgersen           39.5          17.4      186    3800 fema…    2.27
##  3 Adelie  Torgersen           40.3          18        195    3250 fema…    2.24
##  4 Adelie  Torgersen           NA            NA         NA      NA <NA>    NA
##  5 Adelie  Torgersen           36.7          19.3      193    3450 fema…    1.90
##  6 Adelie  Torgersen           39.3          20.6      190    3650 male     1.91
##  7 Adelie  Torgersen           38.9          17.8      181    3625 fema…    2.19
##  8 Adelie  Torgersen           39.2          19.6      195    4675 male     2
##  9 Adelie  Torgersen           34.1          18.1      193    3475 <NA>     1.88
## 10 Adelie  Torgersen           42            20.2      190    4250 <NA>     2.08
## # … with 100 more rows, and abbreviated variable names ¹flipper_length_mm,
## #   ²body_mass_g, ³bill_ratio
## # ℹ Use `print(n = ...)` to see more rows
```

Question 6 (1 pts)

The `slice()` and `slice_min()` functions are useful if we want to select a subset of rows; for example, `slice(1:3)` takes the first three rows, while `slice_min(bill_depth_mm, 3)` takes the three rows with the smallest value of `bill_depth`. This functions also work with `group_by()` so that, for example, `group_by(island) %>% slice(1:3)` takes the first three rows *for each island* (so nine in total).

Take `penguins_2007_br` and, using `group_by` along with either `arrange`, `slice`, or `slice_min`, for *each species* find the three penguins with the shortest bill length. Of those 9 penguins, how many were recorded as female and how many as male?

```
penguins_2007_br %>%
    group_by(species) %>%
    slice_min(bill_length_mm, n = 3)
```

```
## # A tibble: 9 × 8
## # Groups:   species [3]
##   species   island    bill_length_mm bill_depth_mm flipp…¹ body_…² sex   bill_…³
##   <fct>     <fct>              <dbl>         <dbl>   <int>   <int> <fct>   <dbl>
## 1 Adelie    Torgersen           34.1          18.1     193    3475 <NA>     1.88
## 2 Adelie    Torgersen           34.4          18.4     184    3325 fema…    1.87
## 3 Adelie    Torgersen           34.6          21.1     198    4400 male     1.64
## 4 Chinstrap Dream               42.4          17.3     181    3600 fema…    2.45
## 5 Chinstrap Dream               43.2          16.6     187    2900 fema…    2.60
## 6 Chinstrap Dream               45.2          17.8     198    3950 fema…    2.54
## 7 Gentoo    Biscoe              40.9          13.7     214    4650 fema…    2.99
## 8 Gentoo    Biscoe              42            13.5     210    4150 fema…    3.11
## 9 Gentoo    Biscoe              42.8          14.2     209    4700 fema…    3.01
## # … with abbreviated variable names ¹flipper_length_mm, ²body_mass_g,
## #   ³bill_ratio
```

Answer: For the 9 penguins with the shortest `bill_lenght_mm` (3 per species), 7 were recorded as female, 1 was recorded as male, and 1 was recorded as NA.

Question 7 (1 pts)

Using `penguins_2007_br`, calculate the mean and standard deviation of `bill_ratio` for each species using `group_by` and `summarize`. Drop the NAs from `bill_ratio` for these computations (e.g., using the argument `na.rm = TRUE`) so that you have a value for each species. Which species has the greatest average `bill_ratio`?

```
penguins_2007_br %>%
    group_by(species) %>%
    summarize(mean_br = mean(bill_ratio, na.rm = TRUE),
        sd_br = sd(bill_ratio, na.rm = TRUE))
```

```
## # A tibble: 3 × 3
##   species   mean_br sd_br
##   <fct>       <dbl> <dbl>
## 1 Adelie       2.07 0.152
## 2 Chinstrap    2.64 0.169
## 3 Gentoo       3.20 0.157
```

Answer: The Gentoo penguin species has the greatest average `bill_ratio`.

Question 8 (0.5 pts)

With `penguins_2007_br`, using `summarize(n())`, report the number of observations for each species-island combination (note that you'll need to group by both variables!). Which species appears on all three islands?

```
penguins_2007_br %>%
    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
```

Answer: There were 10 Adelie-Biscoe, 20 Adelie-Dream, 20 Adelie-Torgersen, 26 Chinstrap-Dream, and 34 Gentoo-Biscoe observations. The Adelie species appears on all three islands.

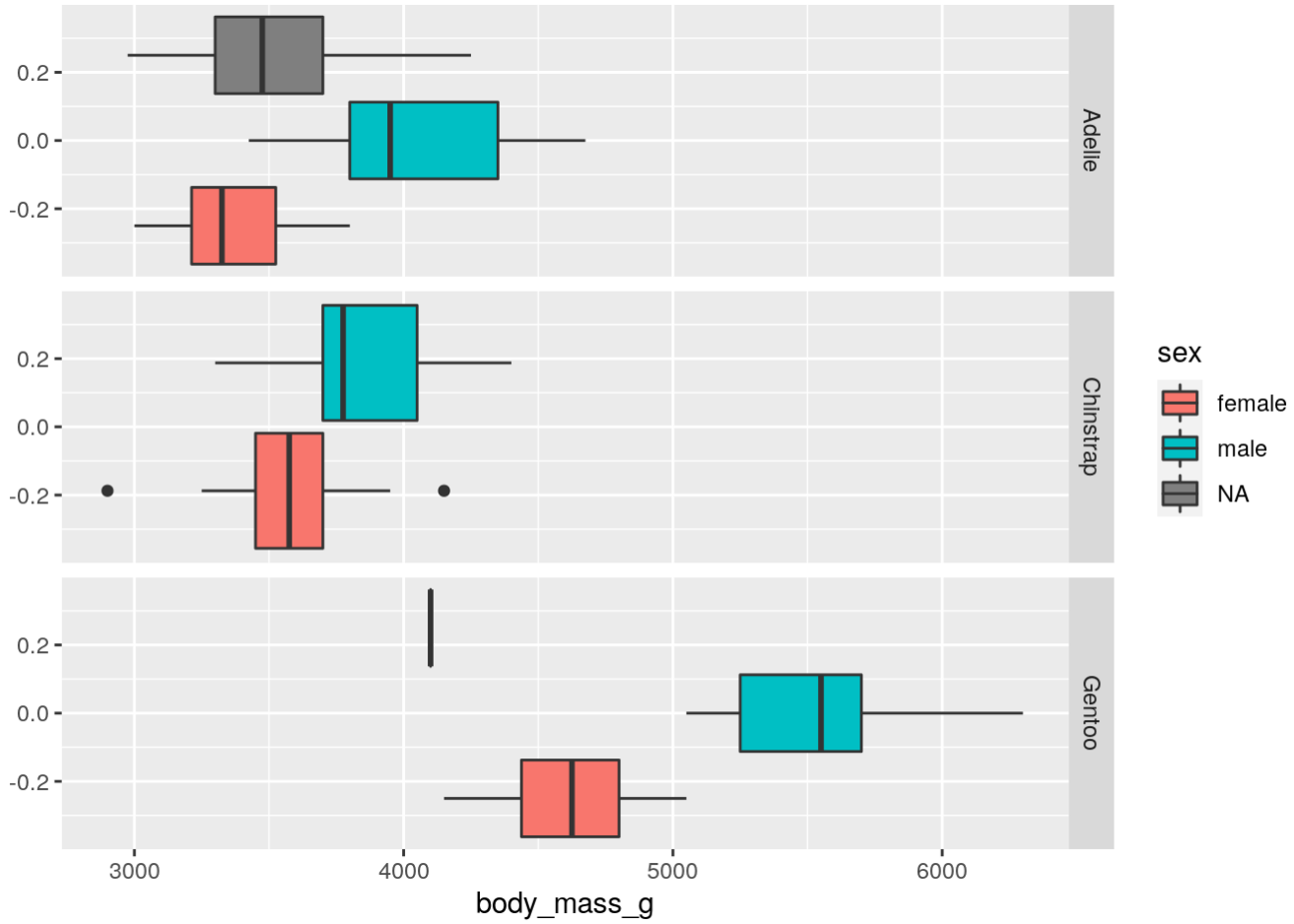
Question 9 (1 pts)

Take the `penguins_2007` data set you have created and using `ggplot`:

1. using whatever `geom`s you think are appropriate, create a single plot showing the distribution of `body_mass_g` for male female penguins separately;

2. facet the plot by species (use `facet_grid` to give each species its own row or column);
3. report below which species (i) has the most NAs for `sex` and (ii) which species shows the least sexual dimorphism (i.e., which shows the greatest overlap of the male/female size distributions).

```
ggplot(penguins_2007, aes(x = body_mass_g, fill = sex)) +
    geom_boxplot() + facet_grid(species ~ .)
```

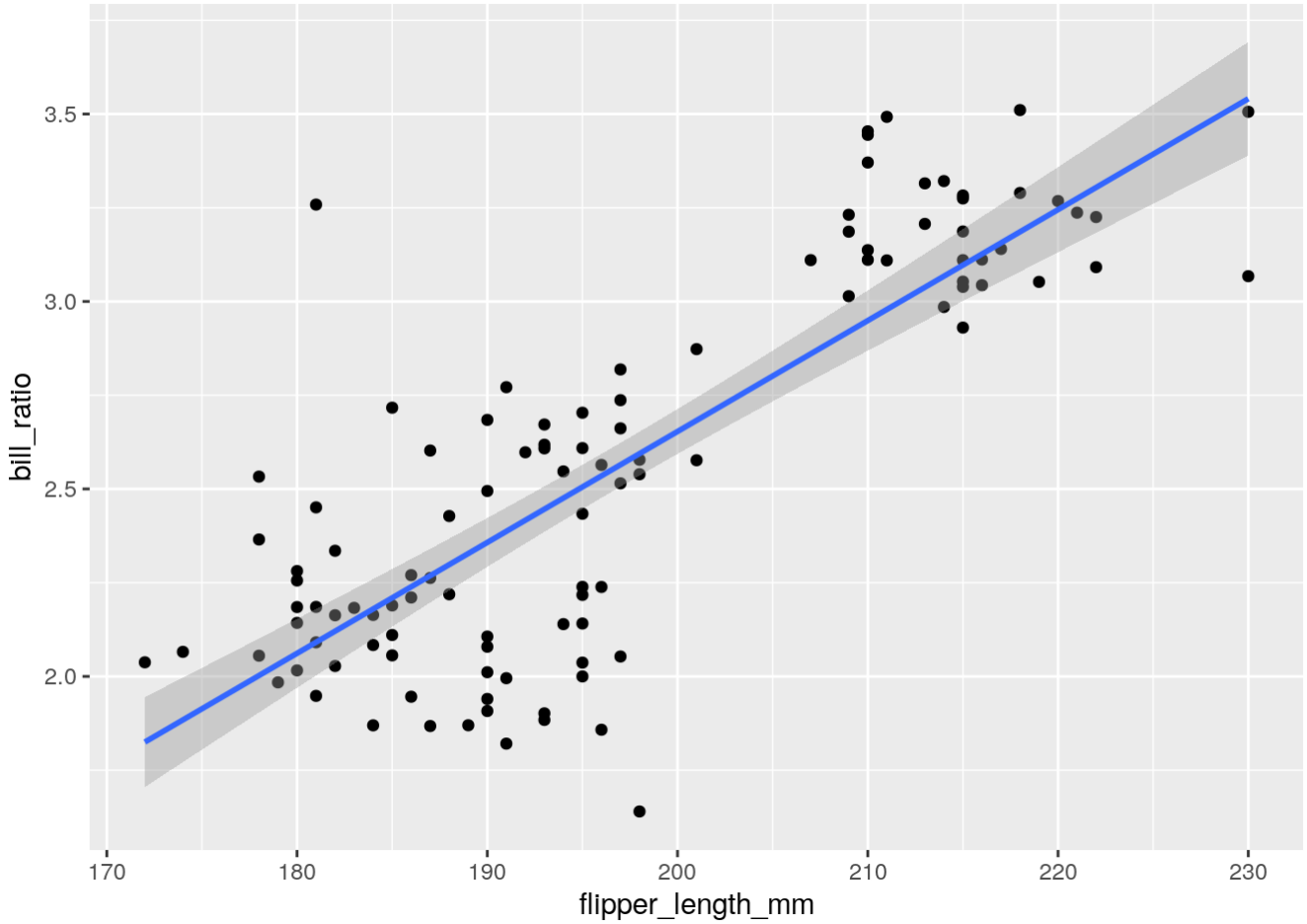


Answer: From the plots above, it can be seen that the Adelie penguins species has the most NAs for `sex`. It can also be seen that the Chinstrap penguin species shows the least sexual dismorphism as they show the greatest overlap of male/female size distributions.

Question 10 (1 pts)

Now, take `penguins_2007_br` and, using `ggplot`, create a scatterplot of `flipper_length_mm` (x-axis) against the `bill_ratio` variable. Does it look like there is a relationship between length-to-depth ratio and the lengths? To see more clearly, add `geom_smooth(method="lm")` to the plot.

```
ggplot(penguins_2007_br, aes(x = flipper_length_mm,
    y = bill_ratio)) + geom_point() + geom_smooth(method = "lm")
```



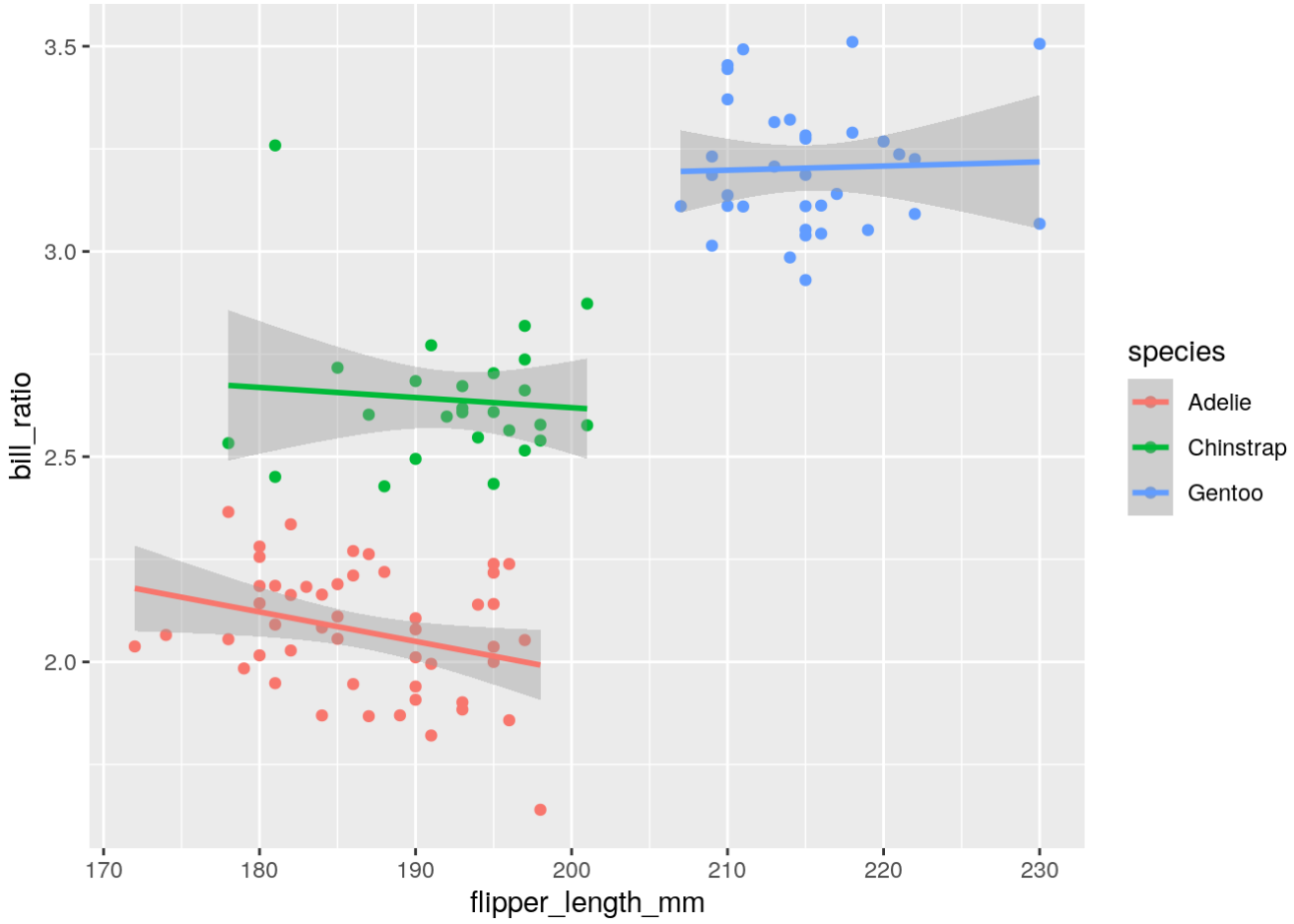
Answer: Overall, there appears to be a relatively positive linear relationship between `flipper_length_mm` and `bill_ratio`.

Question 11 (1 pts)

Does your answer change when you consider each species separately rather than all together? To see more clearly, replicate the plot from the previous question but, additionally, in the main `ggplot()` function map `species` to color so each species gets its own color and smooth.

Compare this plot with the previous one (in 4.2) and discuss whether the relationship between flipper length and bill length-to-depth ratio changes when you look at it overall versus within each species.

```
ggplot(penguins_2007_br, aes(x = flipper_length_mm,
    y = bill_ratio, color = species)) + geom_point() +
    geom_smooth(method = "lm")
```



Answer: When looking at the `flipper_length_mm` to `bill_ratio` relationship per species rather than overall, it becomes apparent that there really isn't much of a relationship between the two properties. The relationship definitely changes when looking at it this way.

Question 12 (1.5 pts)

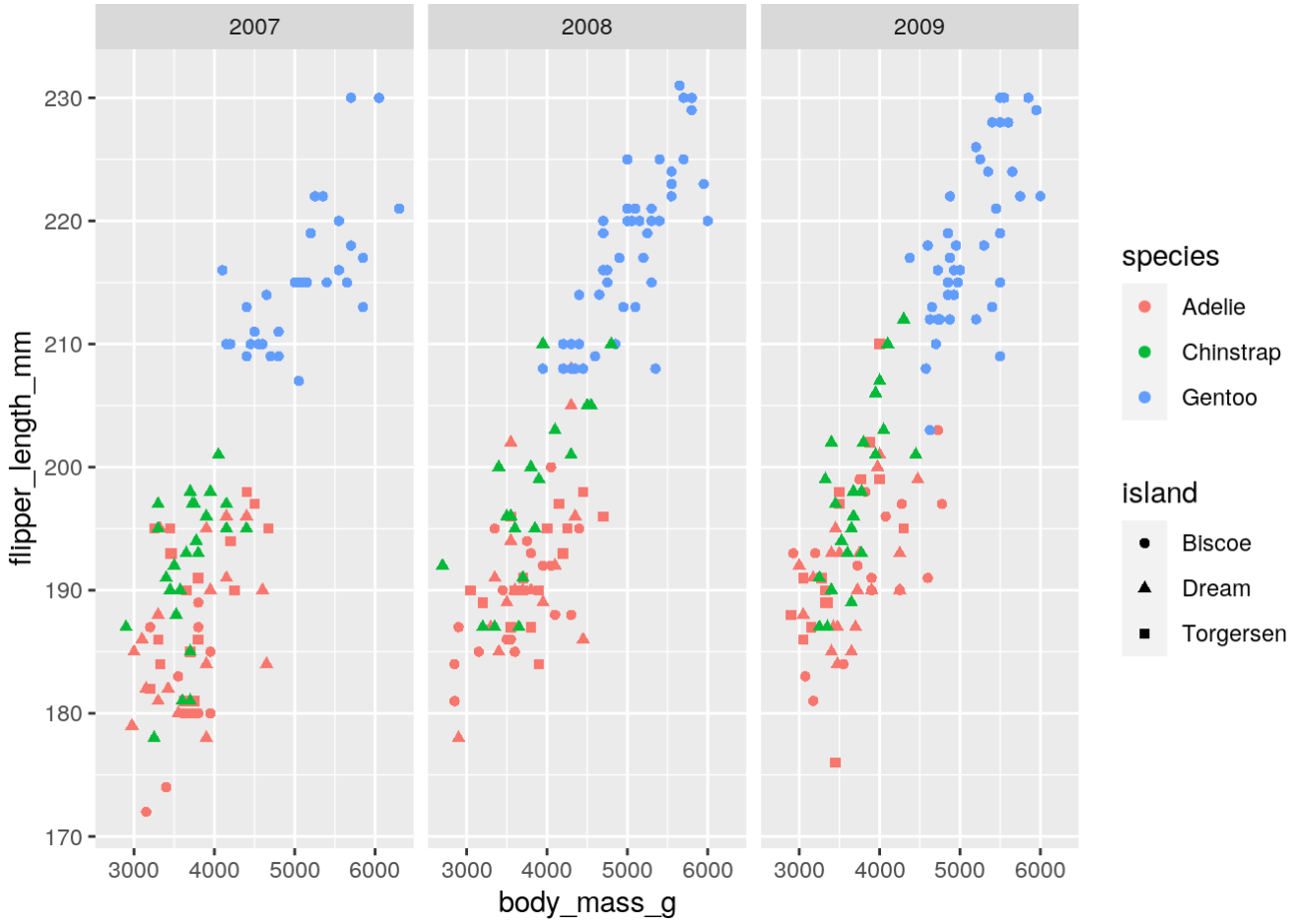
Finally, let's make a plot using the original `penguins` dataset (not just the 2007 data). Forewarning: This will be very busy plot!

Map `body_mass_g` to the x-axis, `flipper_length_mm` to the y-axis, `species` to color, and `island` to shape and make a scatterplot. Using `facet_wrap`, facet the plots by `year`.

Answer the following questions:

1. Does there appear to be a relationship between body mass and flipper length overall?
2. Is there a relationship within each species?
3. What happens to the distribution of flipper lengths for species over time (do you see more or less species overlap for this variable in 2007 relative to 2009)?

```
ggplot(penguins, aes(x = body_mass_g, y = flipper_length_mm,
    color = species, shape = island)) + geom_point() +
    facet_wrap(~year)
```



Answer: Overall, there appears to be a positive linear relationship between `body_mass_g` and `flipper_length_mm`. This same trend also appears to be prevalent within each species (a positive linear relationship). Based on the plots, it appears that the average flipper lengths for Gentoos seems to be relatively the same (maybe increasing slightly on average) and greater than both Chinstraps and Adelies, while the average flipper lengths for Chinstraps and Adelies, which always seem to overlap, appear to be increasing over time as their distributions are slowly starting to overlap with the shorter flipper lengths of Gentoos. Over time, the distribution of flipper lengths for species appears to contain more species overlap in 2009 compared to 2007 as there used to be a more distinct difference in flipper lengths between Gentoos and the other two species, but that is slowly starting to change with the increasing species overlap for this variable.

```
## R version 4.0.3 (2020-10-10)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 18.04.6 LTS
##
## Matrix products: default
## BLAS:   /stor/system/opt/R/R-4.0.3/lib/R/lib/libRblas.so
## LAPACK: /stor/system/opt/R/R-4.0.3/lib/R/lib/libRlapack.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] palmerpenguins_0.1.0 forcats_0.5.1        stringr_1.4.0
##  [4] dplyr_1.0.9          purrr_0.3.4          readr_2.1.2
##  [7] tidyr_1.2.0          tibble_3.1.8         ggplot2_3.3.6
## [10] tidyverse_1.3.2
##
## loaded via a namespace (and not attached):
##  [1] lattice_0.20-45     lubridate_1.8.0     assertthat_0.2.1
##  [4] digest_0.6.29       utf8_1.2.2          R6_2.5.1
##  [7] cellranger_1.1.0    backports_1.4.1     reprex_2.0.1
## [10] evaluate_0.15       highr_0.9           httr_1.4.3
## [13] pillar_1.8.0        rlang_1.0.4         googlesheets4_1.0.0
## [16] readxl_1.4.0        rstudioapi_0.13     jquerylib_0.1.4
## [19] Matrix_1.4-1        rmarkdown_2.14      splines_4.0.3
## [22] labeling_0.4.2      googledrive_2.0.0   munsell_0.5.0
## [25] broom_1.0.0         compiler_4.0.3      modelr_0.1.8
## [28] xfun_0.31           pkgconfig_2.0.3     mgcv_1.8-40
## [31] htmltools_0.5.3     tidyselect_1.1.2    fansi_1.0.3
## [34] crayon_1.5.1        tzdb_0.3.0          dbplyr_2.2.1
## [37] withr_2.5.0         grid_4.0.3          nlme_3.1-158
## [40] jsonlite_1.8.0      gtable_0.3.0        lifecycle_1.0.1
## [43] DBI_1.1.3           magrittr_2.0.3      formatR_1.12
## [46] scales_1.2.0        cli_3.3.0           stringi_1.7.8
## [49] cachem_1.0.6        farver_2.1.1        fs_1.5.2
## [52] xml2_1.3.3          bslib_0.4.0         ellipsis_0.3.2
## [55] generics_0.1.3      vctrs_0.4.1         tools_4.0.3
## [58] glue_1.6.2          hms_1.1.1           fastmap_1.1.0
## [61] yaml_2.3.5          colorspace_2.0-3    gargle_1.2.0
## [64] rvest_1.0.2         knitr_1.39          haven_2.5.0
## [67] sass_0.4.2
```

```
## [1] "2022-09-28 14:52:02 CDT"
```

```
##                                       sysname
##                                       "Linux"
##                                       release
##                             "4.15.0-193-generic"
##                                       version
## "#204-Ubuntu SMP Fri Aug 26 19:20:21 UTC 2022"
##                                      nodename
##                     "educcomp02.ccbb.utexas.edu"
##                                       machine
##                                      "x86_64"
##                                         login
##                                     "unknown"
##                                          user
##                                     "ehc586"
##                                 effective_user
##                                       "ehc586"
```