Data Transformation and Reshaping data

Author: Melina Padron

Question 1:

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.

```
# create a new object with data only from 2007
penguins_2007 <- penguins %>%
  filter(year == 2007)

# count observations from original dataset
nrow(penguins)
```

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

```
# count observations from the new object
nrow(penguins_2007)
```

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

Answer: After creating a new object that only includes observations from 2007 in the Penguin dataset, we can compare the number of observations from 2007 to the entire dataset. There are 344 observations/rows in the original dataset, while there are 110 observations/rows in the new penguins_2007 dataset.

Question 2:

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?

```
# count the number of observations for each species-island combination
penguins_2007 %>%
  group_by(species, island) %>%
  summarize(num_obs = n())
```

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

Answer: After piping the penguin_2007 dataset using group_by() and summarize(), I found that the species that appears on all three islands is Adelie.

Question 3:

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.

```
# added bill_ratio as a new variable to the dataset
penguins_2007 %>%
  mutate(bill_ratio = bill_length_mm / bill_depth_mm)
```

```
##    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 ]
```

```
# overwrote penguin_2007 dataset after checking the new variable was added correctly
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?

```
# filtered new dataset to see if there are observations of bill_ratio 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
```

Answer: After filtering the dataset, I found that there cases in the penguins_2007 dataset for which the `bill_ratio` exceeds 3.5. This is true for the Gentoo species of penguins.

Question 4:

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?

```
# found the three penguins with the smallest bill ratio
penguins_2007 %>%
  group_by(species) %>%
  arrange(bill_ratio) %>%
  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
```

Answer: After grouping the data by species, I arranged the data by bill_ratio to find the three penguins with the smallest bill ratio for each species. The same sex does not have the smallest bill ratio across species. In fact, in the Adelie species the three peguins with the smallest bill ratio came from males, while in the Chinstrap species the three penguins with the smallest bill ratio came from females, and in the Gentoo species the three penguins with the smallest bill ratio came from two females and one male.

Question 5:

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`?

```
# calculated the mean and standard deviation of bill_ratio for each species
# found the species with the greatest mean bill_ratio
penguins_2007 %>%
  group_by(species) %>%
  summarize(mean_bill_ratio = mean(bill_ratio, na.rm = TRUE),
            sd_bill_ratio = sd(bill_ratio, na.rm = TRUE)) %>%
  arrange(desc(mean_bill_ratio)) %>%
  slice(n = 1)
```

```
## # A tibble: 1 x 3
##   species mean_bill_ratio sd_bill_ratio
##   <fct>             <dbl>         <dbl>
## 1 Gentoo             3.20         0.157
```

Answer: After grouping the data by species, I used summarize() to find the mean and standard deviation of bill ratio. I then found that the species with the greatest mean bill ratio was Gentoo.

Question 6:

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)?

```r
# created a plot for each species to show the distribution of body mass by sex
penguins_2007 %>%
  filter(!is.na(sex)) %>%
  ggplot(aes(x = body_mass_g, fill = sex)) +
  geom_histogram() +
  facet_wrap(vars(species), nrow = 3)+
  labs(title = "Distribution of Body Mass by Sex Across Species",
       x = "Body Mass (g)", y = "Frequency")
```

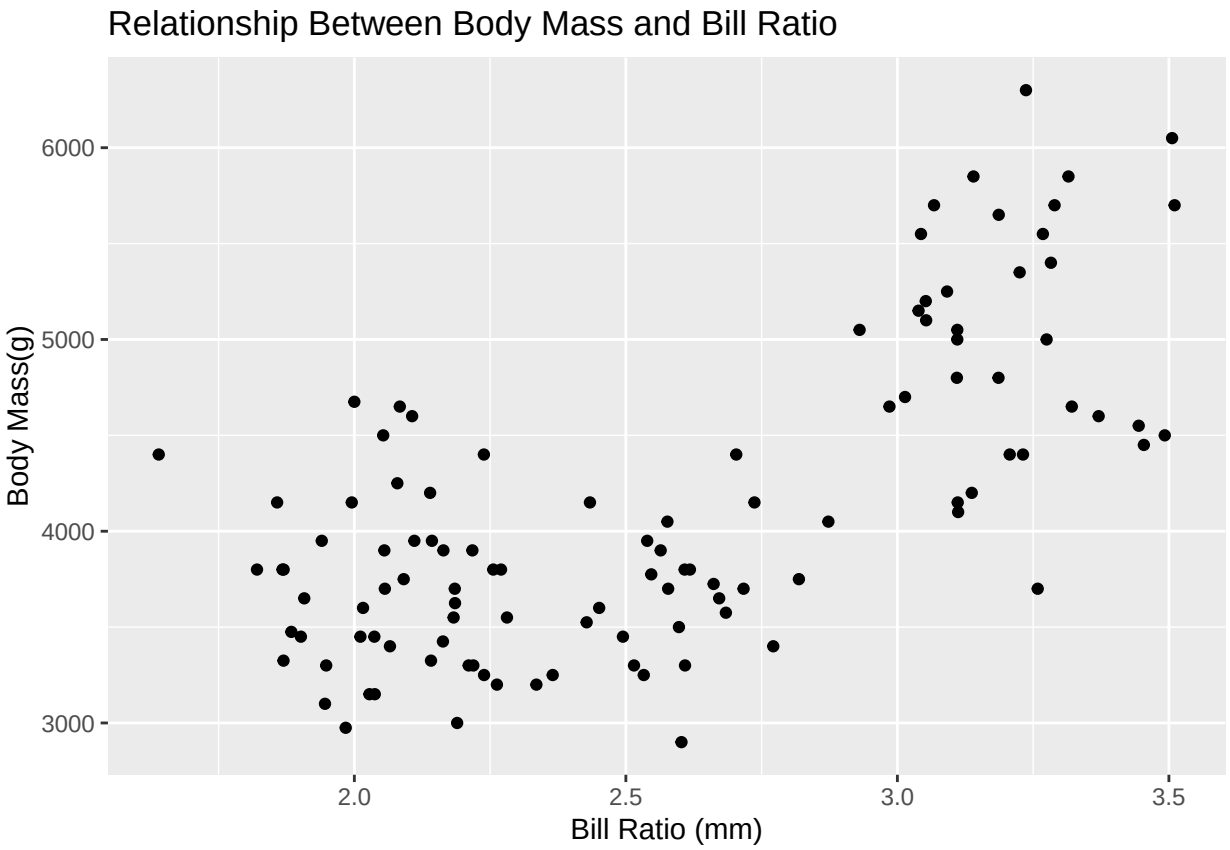
Answer: After creating a plot showing the distribution of body mass by sex for each species, I found that the Chinstrap species shows the least sexual dimorphism. In other words, the Chinstrap species shows the greatest overlap of male/female size distributions.

Question 7:

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
# created a scatterplot of body mass (y-axis) against bill ratio (x-axis)
penguins_2007 %>%
  ggplot(aes(x = bill_ratio, y = body_mass_g)) +
```

```
  geom_point() +
  labs(title = "Relationship Between Body Mass and Bill Ratio",
       x = "Bill Ratio (mm)", y = "Body Mass(g)")
```



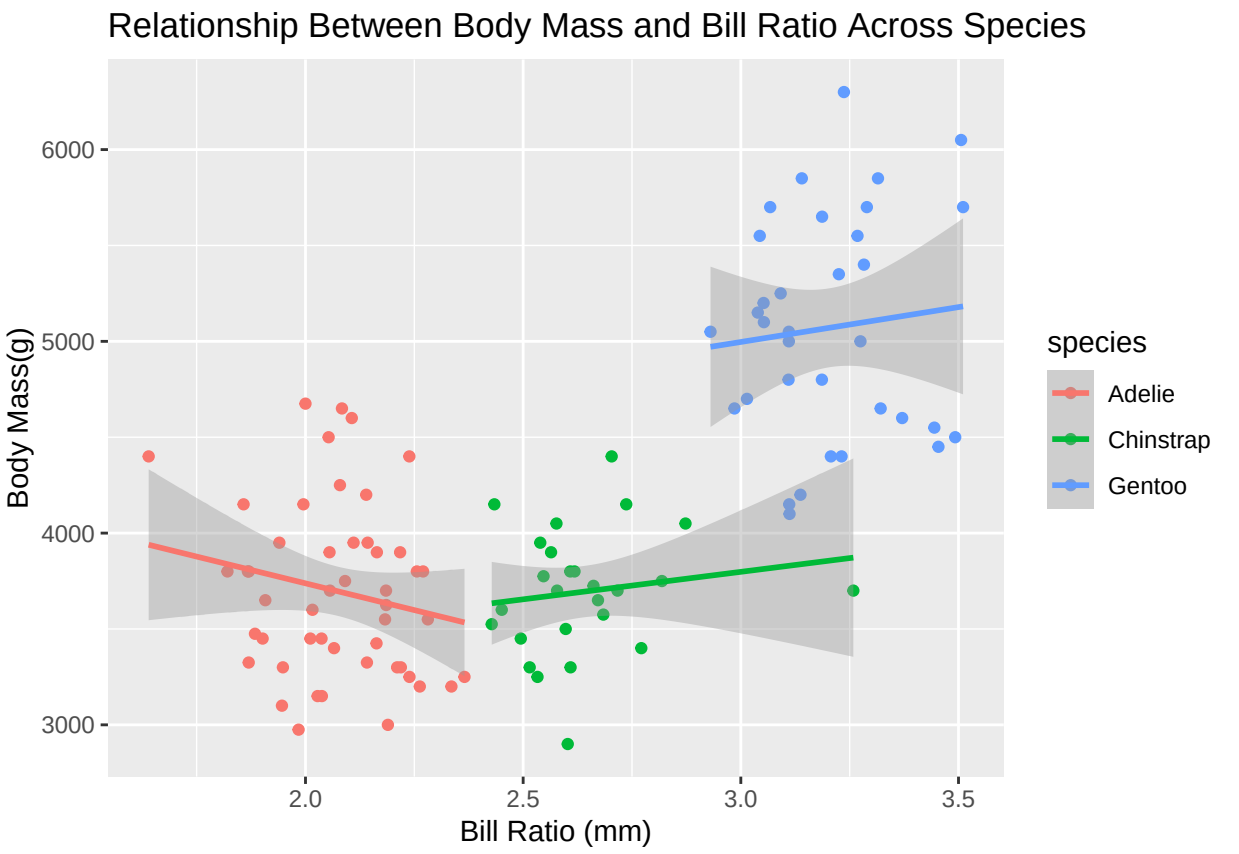
Answer: After creating a scatter plot of body mass against bill ratio, I can see that there is a slightly positive correlation between body mass and bill ratio. This means that on average larger penguins tend to have a larger bill ratio. While creating the scatterplot, I received a warning message that appeared because there were two rows in the penguins_2007 dataset that had missing values for either bill_ratio or body_mass_g, thus those rows were removed from the plot.

Question 8:

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?

```
# created a scatterplot for each species of body mass (y-axis) against bill ratio (x-axis)
# added regression lines
penguins_2007 %>%
  ggplot(aes(x = bill_ratio, y = body_mass_g, color = species)) +
```

```
geom_point() +
geom_smooth(method = "lm", aes(color = species)) +
labs(title = "Relationship Between Body Mass and Bill Ratio Across Species",
     x = "Bill Ratio (mm)", y = "Body Mass(g)")
```



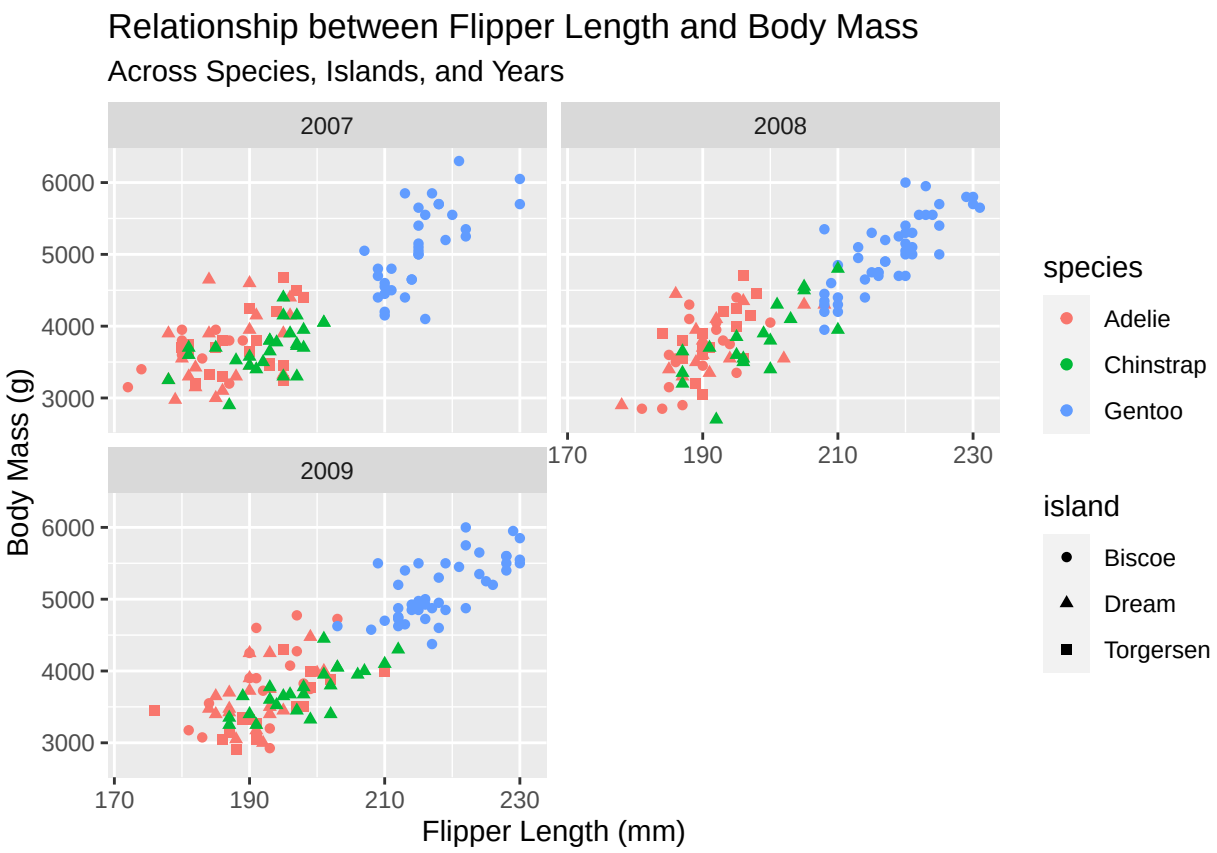
Answer: After creating scatterplots of body mass against bill ratio for each species, I have found that the relationship between the bill ratio and the body mass do not appear to be the same across the different species. Moreover, there appears to be a slightly positive correlation between bill ratio and body mass for the Chinstrap and Gentoo species. On the other hand, there appears to be a negative correlation between bill ratio and body mass for the Adelie species.

Question 9:

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?

```r
# created a scatterplot of flipper length against body mass for each year
# mapped color to species and shape to island
penguins %>%
  ggplot(aes(x = flipper_length_mm, y = body_mass_g, color = species, shape = island)) +
  geom_point() +
  facet_wrap(vars(year), nrow=2) +
  scale_x_continuous(breaks = seq(170, 240, 20)) +
  labs(title = "Relationship between Flipper Length and Body Mass",
       subtitle = "Across Species, Islands, and Years"
       ,x = "Flipper Length (mm)",
       y = "Body Mass (g)")
```



Answer: After creating a scatterplot for each year that compared body mass against flipper length across the different species and islands, I found that there appears to be a positive correlation between body mass and flipper length overall as seen by the positive slope of the different shapes and colors. Moreover, there appears to be a relationship within each species. The Gentoo species tends to have higher body mass and larger flipper lengths while the Chinstrap and Adelie appears to have the roughly the same body mass and flipper length values, with the Chinstrap species having slightly higher values. Over time, we can see that the distribution of higher flipper lengths for species has increased. This means that all species have experienced a growth in flipper lengths over time.
