Code ▾

# R Notebook - Jenna Leali Project 1

# 1.

# Create a R Notebook and install, at least, the following packages: tidyverse, ggplot2, skimr

I already had these packages download on rstudio from a previous class.

# 2.

# Data Loading and Exploration (15 pts):

## a. Load the packages and load the Palmer Penguins dataset.

Hide

```
library(tidyverse)
library(ggplot2)
library(skimr)
palmer <- read.csv("~/Downloads/palmerpenguins (1).csv")
View(palmer)
```

Data Import:

palmer <- read.csv(“~/Downloads/palmerpenguins (1).csv”): Reads a CSV file located at “~/Downloads/palmerpenguins (1).csv” on my computer into a data frame I named “palmer”

Data Viewing:

View(palmer): Opens a data viewer window in RStudio

## b. Display the first few rows of the dataset to examine its structure.

Hide

```
head(palmer, n=10)
```

| | species | island | bill_length_mm | bill_depth_mm | flipper_length_mm | body_mas... | se |
|---|---|---|---|---|---|---|---|
| | <chr> | <chr> | <dbl> | <dbl> | <dbl> | <dbl> | <cl |
| 1 | Adelie | Torgersen | 39.1 | 18.7 | 181 | 3750 | ma |
| 2 | Adelie | Torgersen | 39.5 | 17.4 | 186 | 3800 | fen |
| 3 | Adelie | Torgersen | 40.3 | 18.0 | 195 | 3250 | fen |
| 4 | Adelie | Torgersen | NA | NA | NA | NA | |
| 5 | Adelie | Torgersen | 36.7 | 19.3 | 193 | 3450 | fen |
| 6 | Adelie | Torgersen | 39.3 | 20.6 | 190 | 3650 | ma |
| 7 | Adelie | Torgersen | 38.9 | 17.8 | 181 | 3625 | fen |
| 8 | Adelie | Torgersen | 39.2 | 19.6 | 195 | 4675 | ma |
| 9 | Adelie | Torgersen | 34.1 | 18.1 | 193 | 3475 | |
| 10 | Adelie | Torgersen | 42.0 | 20.2 | 190 | 4250 | |

1-10 of 10 rows

Examine data:

The head() function in R is used to display the first few rows of a data frame. n=10 specifies that I want to look at the first 10 rows.

## c. Provide a brief description of the dataset’s variables

- Species: indicates the species of penguin. For n=10, all the entries belong to the “Adelie” species
- Island: indicates the island the penguin lives on. For n=10, all the entries belong to the “Torgersen” Island.
- Bill_length_mm: represents the bill length of the penguins in millimeters.
- Bill_depth_mm: represents the bill depth of the penguins in millimeters.
- Flipper_length_mm: represents the flipper length of the penguins in millimeters.
- Body_mass_g: represents the body mass of the penguins in grams.
- Sex: indicates the sex of the penguins.
- Year: represents the year in which the data was recorded.

# 3.

# Data Summarization (30 pts):

## a. Calculate summary statistics (mean, median, standard deviation, max, and min) for each relevant numeric variable

Hide

```
summary_statistics <- summary(palmer[, c("bill_length_mm", "bill_depth_mm", "flipper_
length_mm", "body_mass_g")])
print(summary_statistics)
```

```
 bill_length_mm  bill_depth_mm   flipper_length_mm
 Min.   :32.10   Min.   :13.10   Min.   :172.0    
 1st Qu.:39.23   1st Qu.:15.60   1st Qu.:190.0    
 Median :44.45   Median :17.30   Median :197.0    
 Mean   :43.92   Mean   :17.15   Mean   :200.9    
 3rd Qu.:48.50   3rd Qu.:18.70   3rd Qu.:213.0    
 Max.   :59.60   Max.   :21.50   Max.   :231.0    
 NA's   :2       NA's   :2       NA's   :2        
  body_mass_g  
 Min.   :2700  
 1st Qu.:3550  
 Median :4050  
 Mean   :4202  
 3rd Qu.:4750  
 Max.   :6300  
 NA's   :2     
```

Explanation of code:

- The code selects the specific columns “bill_length_mm,” “bill_depth_mm,” “flipper_length_mm,” and “body_mass_g,” from the “palmer” data frame.
- I then used the summary() function to compute summary statistics for each of the selected columns.
- The summary statistics include minimum, 1st quartile, median, mean, 3rd quartile, and maximum values for each of the selected columns.

Explanation of results:

- Bill Length: The bill length ranges from a minimum of 32.10 mm to a maximum of 59.60 mm, with a mean of approximately 43.92 mm. The middle 50% of the data falls within the interquartile range (IQR) of 39.23 mm to 48.50 mm.
- Bill Depth: Similarly, the bill depth varies from 13.10 mm to 21.50 mm, with an average of around 17.15 mm. The IQR for bill depth goes from 15.60 mm to 18.70 mm.

- Flipper Length: Flipper length spans from 172.0 mm to 231.0 mm, with a mean of approximately 200.9 mm. The IQR for flipper length is from 190.0 mm to 213.0 mm.
- Body Mass: Penguin body mass varies from a minimum of 2700 g to a maximum of 6300 g, with an average body mass of roughly 4202 g. The IQR for body mass goes from 3550 g to 4750 g.

## b. Create grouped summaries based on penguin species using the `group_by` and `summarize` functions.

Hide

```
palmer1 <- palmer%>% group_by(species)
summary_statistics <- palmer1 %>%
  summarize(
    mean_bill_length = mean(bill_length_mm, na.rm = TRUE), # Calculate mean bill leng
th
    sd_bill_length = sd(bill_length_mm, na.rm = TRUE),     # Calculate standard devia
tion of bill length
    mean_body_mass = mean(body_mass_g, na.rm = TRUE),     # Calculate mean body mass
    sd_body_mass = sd(body_mass_g, na.rm = TRUE)          # Calculate standard deviat
ion of body mass
  )
print(summary_statistics)
```

| species | mean_bill_length | sd_bill_length | mean_body_mass | sd_body_mass |
|---|---|---|---|---|
| <chr> | <dbl> | <dbl> | <dbl> | <dbl> |
| Adelie | 38.79139 | 2.663405 | 3700.662 | 458.5661 |
| Chinstrap | 48.83382 | 3.339256 | 3733.088 | 384.3351 |
| Gentoo | 47.50488 | 3.081857 | 5076.016 | 504.1162 |

3 rows

Explanation of code:

- First the code takes the palmer data and groups it based on the “species” column. This creates separate groups for each unique penguin species in the dataset.
- na.rm = TRUE: This setting tells R to remove any missing values in the dataset when performing the calculation.
- “mean_bill_length = mean(bill_length_mm”, calculates and assigns the mean value of the “bill_length_mm” variable to the variable mean_bill_length - similarly this is done for standard deviation of bill length, mean body mass, and standard deviation of body mass.

- The code then prints the summary statistics table, showing the mean and standard deviation for bill length and body mass for each penguin species group.

Explanation of results:

- Adelie penguins have the shortest average bill length at approximately 38.79 mm, with a standard deviation of about 2.66 mm, Chinstrap penguins have a longer average bill length at around 48.83 mm, with a slightly higher standard deviation of about 3.34 mm, Gentoo penguins have the longest average bill length among the three species, measuring approximately 47.50 mm, with a standard deviation of around 3.08 mm.
- Adelie penguins have a lower average body mass of approximately 3700.66 grams, with a standard deviation of about 458.57 grams, Chinstrap penguins have a similar average body mass of approximately 3733.09 grams, with a standard deviation of about 384.34 grams, Gentoo penguins are the heaviest among the three species, with an average body mass of approximately 5076.02 grams and a standard deviation of about 504.12 grams.
- Adelie penguins tend to have shorter bills and lower body masses, while Gentoo penguins have longer bills and higher body masses. Chinstrap penguins fall in between these two species in terms of bill length and body mass.

# c. Discuss the insights gained from the summarization process. Note any interesting patterns, anomalies, missing, etc

Species Similarities and Differences:

- Chinstrap and Gentoo penguins have similar mean body masses, Chinstrap penguins have notably longer bills on average. Adelie penguins have a slightly lower mean body mass compared to the other two species, but their bill length is significantly shorter on average.

Bill Length:

- The standard deviation of bill length for Chinstrap penguins is relatively higher (3.34) compared to Adelie (2.66) and Gentoo (3.08) penguins. This means that there might be more variability in bill length within the Chinstrap species compared to the other two.

Body Mass Differences:

- Gentoo penguins have the highest mean body mass at approximately 5076 grams, followed by Chinstrap penguins at 3733 grams, and Adelie penguins at 3701 grams.

Missing Data:

- The summarization results do not show any missing values (NAs) for the calculated summary statistics.

Summary:

- The summary showed significant differences in bill length and body mass among the three penguin species, with Chinstrap penguins having the longest bills and Gentoo penguins having the highest mean body mass.

## d. Apply the skimr package to the data and discuss the output. How does it compare to what you did in the previous steps? What additional information is provided? How is this useful?

### I first create a copy of the original dataset for visualization before applying the skimr package. I had problems making visualizations without doing this first.

Hide

```
original_palmer <- palmer
summary_stats <- skim(palmer)
print(summary_stats)
```

```
── Data Summary ────────────────────────
                           Values
Name                       palmer
Number of rows             344
Number of columns          8
_______________________
Column type frequency:
  character                3
  numeric                  5
________________________
Group variables            None

── Variable type: character ────────────────────────────────────
  skim_variable n_missing complete_rate   min   max empty
1 species               0             1     6     9     0
2 island                0             1     5     9     0
3 sex                   0             1     0     6    11
  n_unique whitespace
1        3          0
2        3          0
3        3          0

── Variable type: numeric ──────────────────────────────────────
  skim_variable     n_missing complete_rate   mean      sd
1 bill_length_mm            2         0.994   43.9   5.46
2 bill_depth_mm             2         0.994   17.2   1.97
3 flipper_length_mm         2         0.994  201.   14.1
4 body_mass_g               2         0.994 4202.  802.
5 year                      0         1     2008.    0.818
      p0    p25    p50    p75   p100 hist
1   32.1   39.2   44.4   48.5   59.6 ▃▇▇▆▁
2   13.1   15.6   17.3   18.7   21.5 ▅▅▇▇▂
3  172    190    197    213    231   ▂▇▃▅▂
4 2700   3550   4050   4750   6300   ▃▇▆▃▂
5 2007   2007   2008   2009   2009   ▇▁▇▁▇
```

Explanation of code:

- I first created a duplicate of the “palmer” dataset and stored it in a new data frame called “original_palmer.” This step preserves the original data for when I make visualizations.
- I then used the skim() function from the skimr package to compute a set of summary statistics for each column in the “palmer” dataset
- I then printed the summary statistics

Explanation of results:

- The dataset contains a total of 344 rows and 8 columns.
- Three columns are character data (“species,” “island,” and “sex”) and five columns are numeric data (bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g, and “year)
- species has 3 unique categories, island also has 3 unique categories, and sex has 3 unique categories
- The numeric variables provide mean and standard deviation.

Comparison with Previous Step:

- With dplyr grouped summaries, the results focused on specific summary statistics (e.g., mean, standard deviation) for numeric variables and grouped the data by a particular variable (species) to compare those statistics among groups.
- With skimr for dataset summary statistics, I obtained a broader overview of your entire dataset, including information on data types, missing values, completeness rates, and a wide range of statistics (e.g., mean, standard deviation, quartiles) for all numeric variables.
- The skimr output is more comprehensive and provides a better understanding of the dataset. Overall, dplyr is more targeted summaries for specific questions about grouped data, while skimr is used for a for an overall understanding of the entire dataset’s characteristics.

I now revert back to the original dataset for visualization

Hide

```
palmer <- original_palmer
```

# 4.

# Data Visualization (30 pts):

## a. Create at least three different types of visualizations using ggplot2 (e.g., scatter plot, bar plot, box plot, histogram, etc.) to explore relationships between variables.

## b. Ensure appropriate labeling, coloring, and titling of the visualizations.

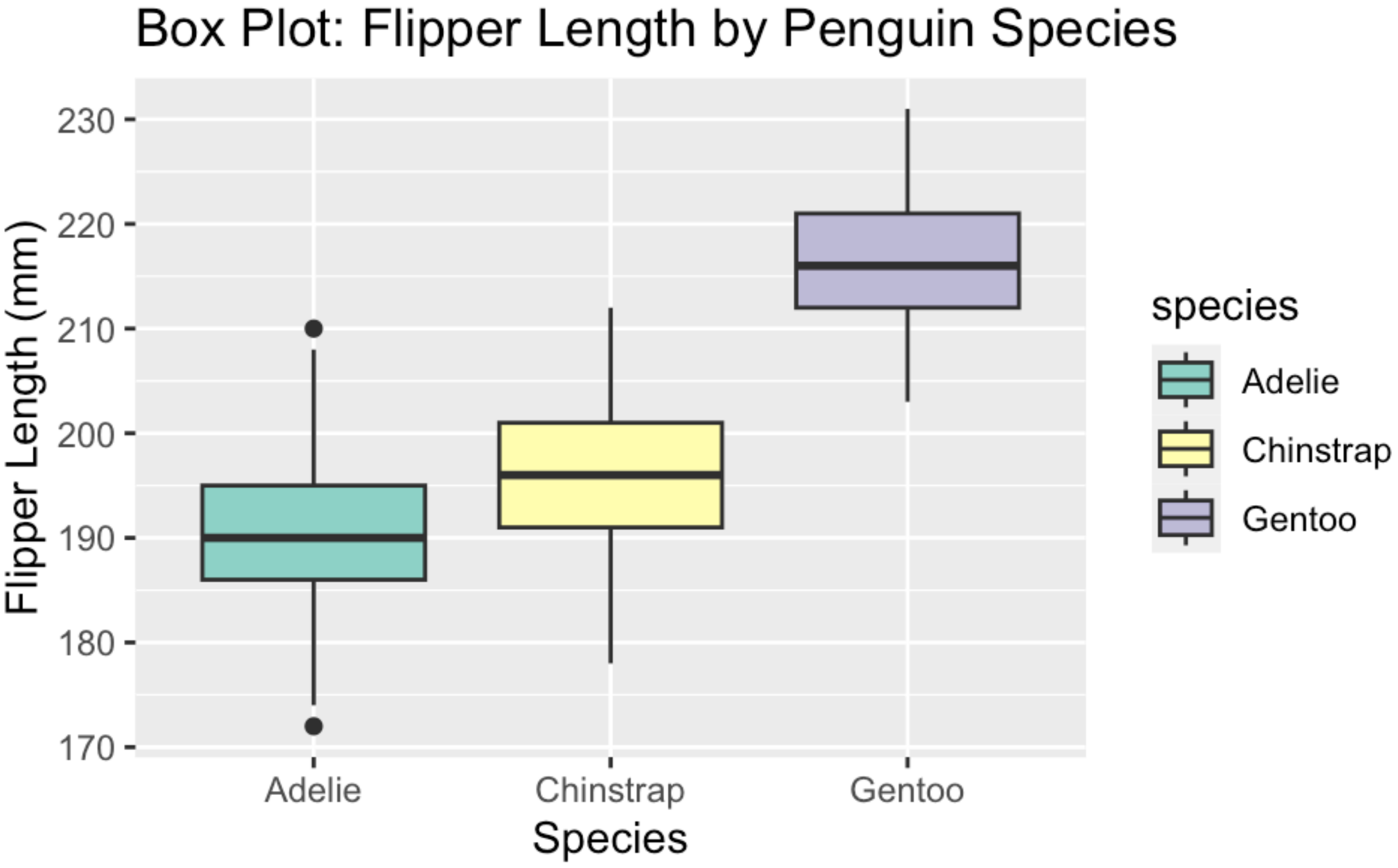
## c. Interpret the insights obtained from each visualization.

Box Plot: Flipper Length by Species

Hide

```
ggplot(data = palmer, aes(x = species, y = flipper_length_mm, fill = species)) +
  geom_boxplot() +
  labs(x = "Species", y = "Flipper Length (mm)") +
  ggtitle("Box Plot: Flipper Length by Penguin Species") +
  scale_fill_brewer(palette = "Set3")
```

```
Warning: Removed 2 rows containing non-finite values (`stat_boxplot()`).
```



Insights:

- The first box represents the Adelie penguin species. The line inside the box is located at 190 mm, which is the median flipper length for Adelie penguins. The box itself spans from approximately 186 mm to 195 mm. The lines of the box plot extend from the box to the minimum value of around 174 mm and the maximum value of approximately 207 mm. There are two dots outside the lines, one at 172 mm and another at 210 mm, which represent potential outliers.
- The second box corresponds to the Chinstrap penguin species. The line inside the box is located at 196 mm, which is the median flipper length for Chinstrap penguins.The box spans from approximately 191 mm to 201 mm. The lines extend from the box to the minimum value of 187 mm and the maximum value of 212 mm. There are no dots outside the lines, indicating no apparent outliers.

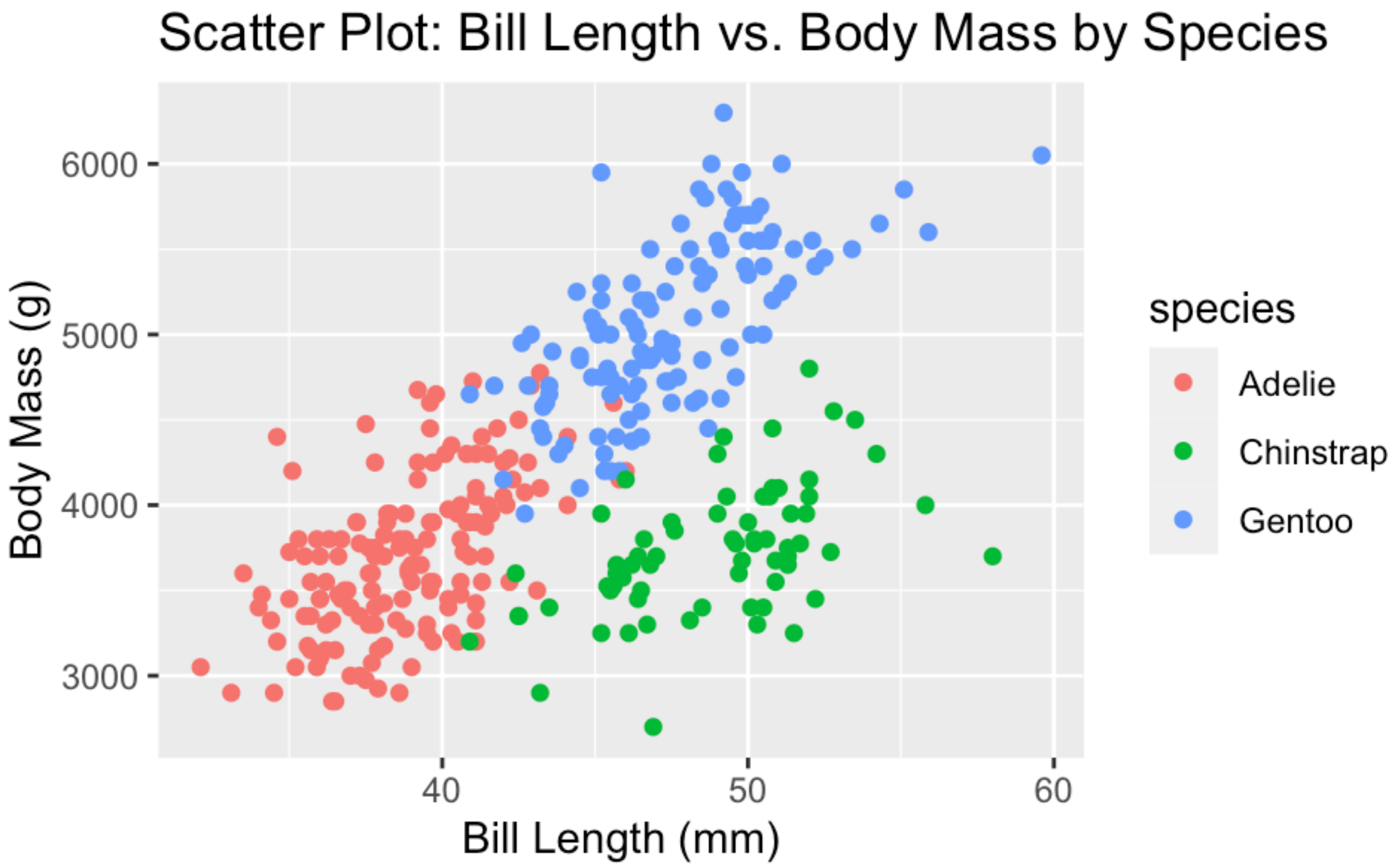
- The third box represents the Gentoo penguin species.The line inside the box is located at 216 mm, which is the median flipper length for Gentoo penguins. The box spans from approximately 213 mm to 221 mm. The lines extend from the box to the minimum value of 203 mm and the maximum value of 231 mm. There are no dots outside the lines, indicating no apparent outliers.
- Overall, the Gentoo penguins tend to have the longest flipper lengths, followed by the Chinstrap penguins, while the Adelie penguins have the shortest flipper lengths. The presence of outliers in the Adelie species means some variability or potential extreme values in their flipper lengths.

Scatter Plot: Bill Length vs. Body Mass

Hide

```
ggplot(data = palmer, aes(x = bill_length_mm, y = body_mass_g, color = species)) +
  geom_point() +
  labs(x = "Bill Length (mm)", y = "Body Mass (g)") +
  ggtitle("Scatter Plot: Bill Length vs. Body Mass by Species")
```

```
Warning: Removed 2 rows containing missing values (`geom_point()`).
```



Insights:

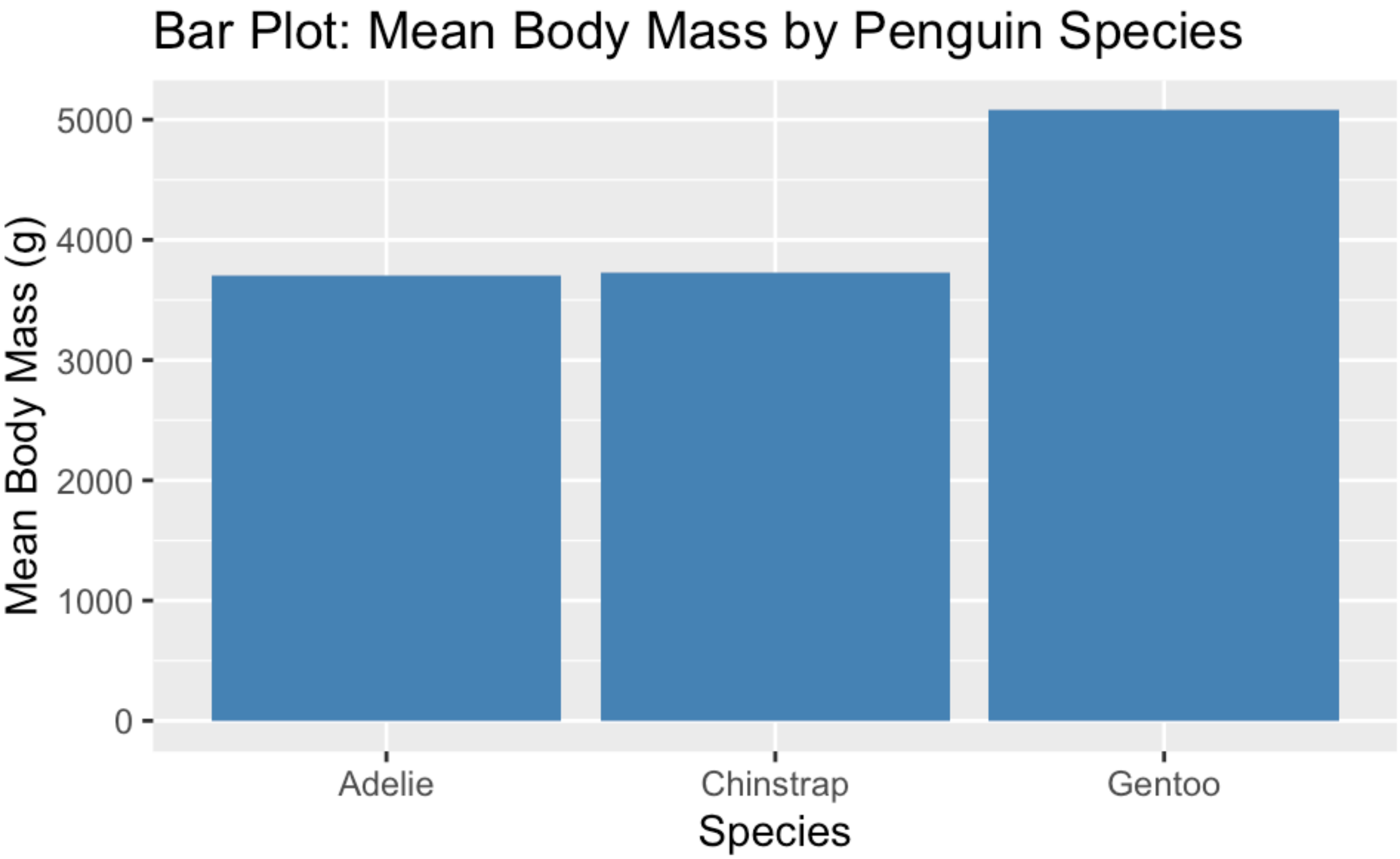
- Adelie Species (Red Points): The scatterplot shows that for the Adelie penguin species, body mass varies across a range of approximately 2900 to 4600 grams. The majority of Adelie penguins have a body mass that falls within the range of 3000 to 4000 grams, as indicated by the dense clustering of red points in this region. Regarding bill length, Adelie penguins exhibit a wide range, with bill lengths spanning from 5 to 46 millimeters. The majority of Adelie penguins have bill lengths that cluster between 20 and 43 millimeters, with the highest concentration of points in this range.
- Chinstrap Species (Green Points): For the Chinstrap penguin species, body mass varies across a range of approximately 2700 to 4800 grams. Most Chinstrap penguins have a body mass that falls within the range of 3400 to 4100 grams, as indicated by the dense clustering of green points in this region. In terms of bill length, Chinstrap penguins have bill lengths that range from 40 to 57 millimeters. The majority of Chinstrap penguins have bill lengths clustered between 45 and 53 millimeters, with a higher concentration of red points in this range.
- Gentoo Species (Blue Points): For the Gentoo penguin species, body mass varies across a wider range, from approximately 4000 to 6300 grams. The majority of Gentoo penguins have a body mass between 4500 and 5700 grams, as evidenced by the clustering of blue points in this range. In terms of bill length, Gentoo penguins have bill lengths ranging from 41 to 59 millimeters. The majority of Gentoo penguins have bill lengths clustered between 45 and 53 millimeters, with a high concentration of green points in this range.
- Overall, in all three penguin species, there appears to be a positive correlation between bill length and body mass. This means that as bill length increases, body mass tends to increase as well. This positive correlation is evident by the general trend of points sloping upwards from left to right in each species cluster. While there is a general positive correlation, it’s important to note that there is variability within each species. This variability can be observed by the spread of points around the trendline.

Bar Plot: Mean Body Mass by Species

Hide

```
ggplot(data = palmer, aes(x = species, y = body_mass_g)) +
       geom_bar(stat = "summary", fun = "mean", fill = "steelblue") +
    labs(x = "Species", y = "Mean Body Mass (g)") +     ggtitle("Bar Plot: Mean Body
Mass by Penguin Species")
```

```
Warning: Removed 2 rows containing non-finite values (`stat_summary()`).
```



Insights:

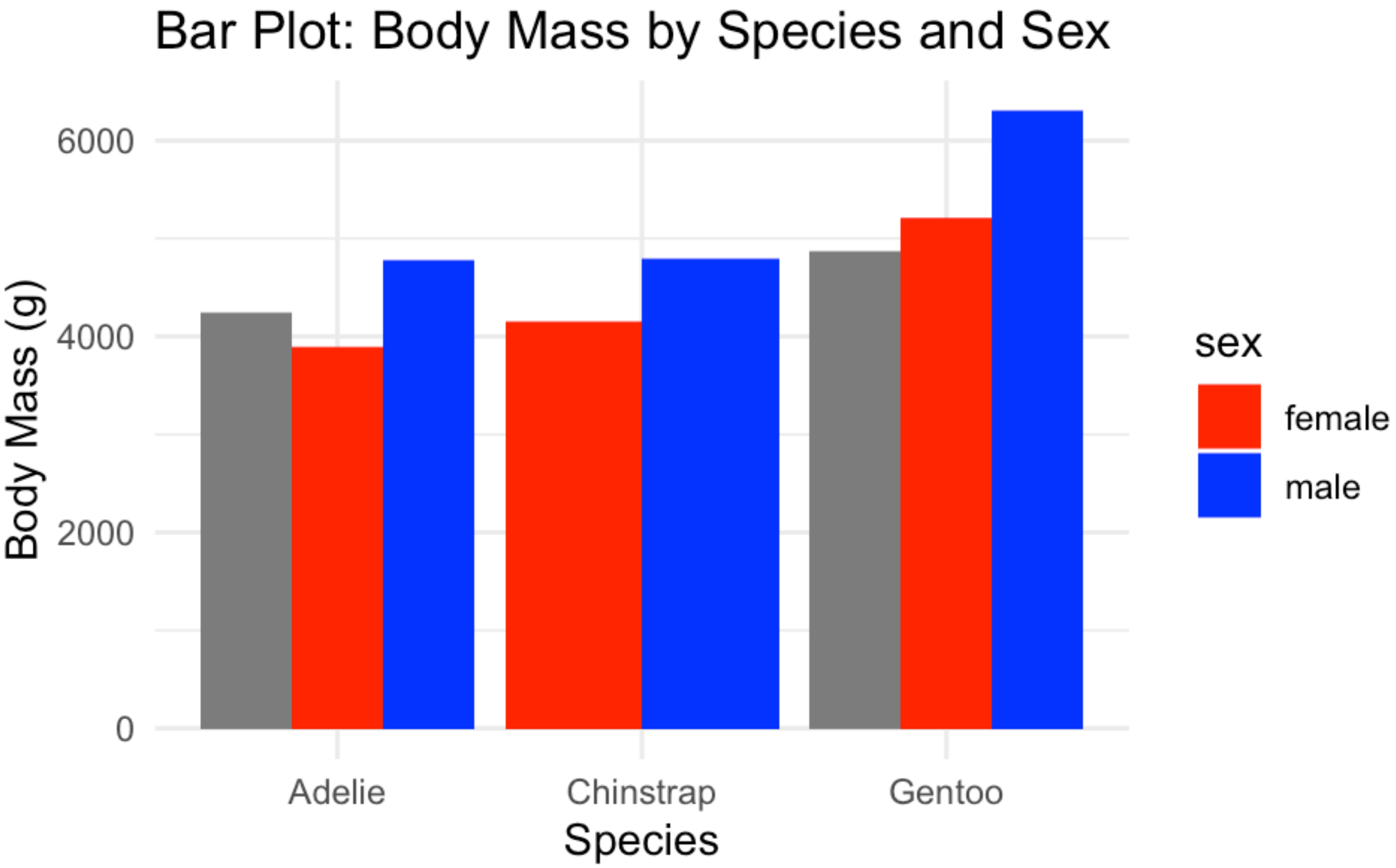
- The first bar represents the Adelie penguin species. The bar reaches up to approximately 3700 grams on the y-axis. This means that, on average, Adelie penguins have a mean body mass of around 3700 grams.
- The second bar represents the Chinstrap penguin species. The bar extends to approximately 3750 grams on the y-axis. This indicates that, on average, Chinstrap penguins have a mean body mass of approximately 3750 grams.
- The third bar is the Gentoo penguin species. The bar goes up to approximately 5100 grams on the y-axis. This shows that, on average, Gentoo penguins have a significantly higher mean body mass, with an approximate value of 5100 grams.
- The bar plot shows the differences in mean body mass among these three penguin species. Gentoo penguins have the highest average body mass, followed by Chinstrap penguins, while Adelie penguins have the lowest average body mass.

Bar Plot: Body Mass by Species and Sex within Species

Hide

```
ggplot(palmer, aes(x = species, y = body_mass_g, fill = sex)) +
    geom_bar(stat = "identity", position = "dodge") +
    labs(x = "Species", y = "Body Mass (g)") +
    ggtitle("Bar Plot: Body Mass by Species and Sex") +   scale_fill_manual(values =
c("male" = "blue", "female" = "red")) +
    theme_minimal()
```

```
Warning: Removed 2 rows containing missing values (`geom_bar()`).
```



Insights:

- Adelie Species: The grey line corresponds to the “Not Specified” category, which includes rows where the “sex” variable is missing or set to “Not Specified.” This category has a maximum body mass value of 4100 g. The red line represents female Adelie penguins. It has a maximum body mass value of 3900 g. The blue line represents male Adelie penguins. It has a maximum body mass value of 4900 g.
- Chinstrap Species: The red line represents female Chinstrap penguins. It has a maximum body mass value of 4080 g. The blue line represents male Chinstrap penguins. It has a maximum body mass value of 4900 g.

- Gentoo Species: Similar to the Adelie species, the grey line corresponds to the “Not Specified” category, which includes rows where the “sex” variable is missing or set to “Not Specified.” This category has a maximum body mass value of 4950 g. The red line represents female Gentoo penguins. It has a maximum body mass value of 5100 g. The blue line represents male Gentoo penguins. It has a maximum body mass value of 6200 g.
- In all species the male penguins have a higher body mass index than the female penguins. This difference is most noticable in Gentoo penguins, where the blue bar is substantially taller.

# 5.

# Project Report and Interpretation (25 pts):

## a. Compile a comprehensive project report either directly in the R Notebook or, if you decided not to use a notebook format, an R Script plus a Word document.

## b. Summarize overall patterns, trends, or relationships you discovered. What can you say about each penguin?

Adelie Penguins:

- Adelie penguins, can be characterized by their shorter flipper lengths and lower mean body mass, additionally, males have a higher body mass than females. The median flipper length for Adelie penguins is 190 mm, with most flipper lengths falling between approximately 186 mm and 195 mm. In terms of body mass, Adelie penguins show a range of approximately 2900 to 4600 grams, with a majority clustered between 3000 and 4000 grams. The bill length varies widely, spanning from 5 to 46 millimeters, with a concentration between 20 and 43 millimeters. Despite the presence of potential outliers, this species displays a positive correlation between bill length and body mass.

Chinstrap Penguins:

- Chinstrap penguins, with a slightly higher mean body mass than Adelie penguins, exhibit a similar pattern, where males have a higher body mass than females. The median flipper length for Chinstrap penguins is 196 mm, with most flipper lengths concentrated between approximately 191 mm and 201 mm. In terms of body mass, Chinstrap penguins have body masses ranging from approximately 2700 to 4800 grams, with most falling between 3400 and 4100 grams. Bill lengths range from 40 to 57 millimeters, with a concentration between 45 and 53 millimeters. This species does not display apparent outliers and shows a positive correlation between bill length and body mass.

Gentoo Penguins:

- Gentoo penguins distinguish themselves with the longest flipper lengths and the highest mean body mass among the three species. Males have a substantially higher body mass than females. The median flipper length for Gentoo penguins is 216 mm, with most flipper lengths ranging from approximately 213 mm to 221 mm. In terms of body mass, Gentoo penguins have a wide range, from approximately 4000 to 6300 grams, with a majority clustered between 4500 and 5700 grams. Bill lengths range from 41 to 59 millimeters, with a concentration between 45 and 53 millimeters. Gentoo penguins do not show apparent outliers and also display a positive correlation between bill length and body mass.

Comparison:

- Gentoo penguins stand out with the longest flipper lengths and the highest mean body mass, followed by Chinstrap penguins, while Adelie penguins have shorter flipper lengths and the lowest mean body mass. In all three species, males generally have a higher body mass than females. The positive correlation between bill length and body mass is a common trend across the species, suggesting a potential relationship between these two physical characteristics.

Using R and Tidyverse:

- Using R and the tidyverse for the analysis of the penguin data provides many advantages. The tidyverse’s suite of packages, such as ggplot2 and dplyr, offers powerful tools specifically tailored to efficiently handle, clean, and visualize complex datasets like the penguin dataset. With these tools, I was able to easily generate informative visualizations like box plots, scatter plots, and bar plots, which helped uncover patterns and trends in penguin characteristics.