# Exploratory Data Analysis of Penguins Dataset

aurellaa

2025-02-14

# 1. Introduction

The *palmerpenguins* package provides size measurements of three penguin species from the Palmer Archipelago, Antarctica. This project conducts an exploratory data analysis (EDA) on the *penguins* dataset (a simplified version of the raw data) in order to investigate the factors that influence penguin body mass. By identifying predictors of penguin body, this report aims to provide a deeper understanding of penguin morphology and health.

# 2. Data Inspection and Cleaning

## 2.1 Loading Required Libraries

Below is a list of the packages used in this report:

```
library(palmerpenguins)
library(ggplot2)
library(ggpubr)
library(skimr)
library(dplyr)
library(gridExtra)
library(cowplot)
```

## 2.2 Inspecting the Dataset

Inspecting the dataset reveals that there are 344 observations for 8 variables: species, island, bill length (mm), bill depth (mm), flipper length (mm), body mass (g), sex, and year. There are three penguin species, “Adelie”, “Chinstrap”, and “Gentoo”, found across the islands “Biscoe”, “Dream”, and “Torgersen”. The observations were recorded within a span of 3 years, from 2007 to 2009.

```
## 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 ...
```

```
##       species          island    bill_length_mm  bill_depth_mm
##  Adelie   :152   Biscoe   :168   Min.   :32.10   Min.   :13.10
##  Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60
##  Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30
##                                   Mean   :43.92   Mean   :17.15
##                                   3rd Qu.:48.50   3rd Qu.:18.70
##                                   Max.   :59.60   Max.   :21.50
##                                   NA's   :2       NA's   :2
##  flipper_length_mm  body_mass_g       sex           year
##  Min.   :172.0     Min.   :2700   female:165   Min.   :2007
##  1st Qu.:190.0     1st Qu.:3550   male  :168   1st Qu.:2007
##  Median :197.0     Median :4050   NA's  : 11   Median :2008
##  Mean   :200.9     Mean   :4202                Mean   :2008
##  3rd Qu.:213.0     3rd Qu.:4750                3rd Qu.:2009
##  Max.   :231.0     Max.   :6300                Max.   :2009
##  NA's   :2         NA's   :2
```

The uneven number of observations per species should be taken into consideration when interpreting statistical analyses, with Adelie penguins being the most represented (152 individuals) and Chinstrap penguins the least (52 individuals). This imbalance in sample size may influence results.

It should also be noted that there are some missing values, which may have arisen due to incomplete data collection during fieldwork. Environmental factors, for instance, may have posed a difficulty for researchers to measure certain variables consistently. The missing values in sex may be due to challenges in visually distinguishing female and male penguins.

The following shows all of the missing values present:

```
##           species            island    bill_length_mm     bill_depth_mm
##                 0                 0                 2                 2
## flipper_length_mm       body_mass_g               sex              year
##                 2                 2                11                 0
```

## 2.3 Handling Missing Values

Omission of missing values may impose bias on the results of the statistical analyses.

```
penguins_clean <- na.omit(penguins)
```

# 3. Descriptive Statistics

## 3.1 Overall Summary

Summary statistics are computed for the clean data.

```
summary(penguins_clean)
```

```
##       species          island    bill_length_mm  bill_depth_mm
##  Adelie   :146   Biscoe   :163   Min.   :32.10   Min.   :13.10
##  Chinstrap: 68   Dream    :123   1st Qu.:39.50   1st Qu.:15.60
##  Gentoo   :119   Torgersen: 47   Median :44.50   Median :17.30
##                                   Mean   :43.99   Mean   :17.16
##                                   3rd Qu.:48.60   3rd Qu.:18.70
##                                   Max.   :59.60   Max.   :21.50
##  flipper_length_mm  body_mass_g       sex           year
##  Min.   :172       Min.   :2700   female:165   Min.   :2007
##  1st Qu.:190       1st Qu.:3550   male  :168   1st Qu.:2007
##  Median :197       Median :4050                Median :2008
##  Mean   :201       Mean   :4207                Mean   :2008
##  3rd Qu.:213       3rd Qu.:4775                3rd Qu.:2009
##  Max.   :231       Max.   :6300                Max.   :2009
```

```
skim(penguins_clean)
```

Data summary

| | |
|---|---|
| Name | penguins_clean |
| Number of rows | 333 |
| Number of columns | 8 |
| _____ | |
| Column type frequency: | |
| factor | 3 |
| numeric | 5 |
| _____ | |
| Group variables | None |

**Variable type: factor**

| skim_variable | n_missing | complete_rate | ordered | n_unique | top_counts |
|---|---|---|---|---|---|
| species | 0 | 1 | FALSE | 3 | Ade: 146, Gen: 119, Chi: 68 |
| island | 0 | 1 | FALSE | 3 | Bis: 163, Dre: 123, Tor: 47 |
| sex | 0 | 1 | FALSE | 2 | mal: 168, fem: 165 |

**Variable type: numeric**

| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
|---|---|---|---|---|---|---|---|---|---|---|
| bill_length_mm | 0 | 1 | 43.99 | 5.47 | 32.1 | 39.5 | 44.5 | 48.6 | 59.6 | |
| bill_depth_mm | 0 | 1 | 17.16 | 1.97 | 13.1 | 15.6 | 17.3 | 18.7 | 21.5 | |
| flipper_length_mm | 0 | 1 | 200.97 | 14.02 | 172.0 | 190.0 | 197.0 | 213.0 | 231.0 | |
| body_mass_g | 0 | 1 | 4207.06 | 805.22 | 2700.0 | 3550.0 | 4050.0 | 4775.0 | 6300.0 | |
| year | 0 | 1 | 2008.04 | 0.81 | 2007.0 | 2007.0 | 2008.0 | 2009.0 | 2009.0 | |

## 3.2 Summary Statistics by Species

The summary statistics are further broken down by species to examine the morphological differences between Adelie, Chinstrap, and Gentoo penguins. Differences in average body mass, bill length, bill depth, and flipper length may reflect species-specific adaptations to their respective habitats. This comparison helps assess whether species characteristics could be potential predictors of body mass.

```
penguins_clean |>
  group_by(species) |>
  summarize(
    avg_body_mass = mean(body_mass_g),
    avg_bill_length = mean(bill_length_mm),
    avg_bill_depth = mean(bill_depth_mm),
    avg_flipper_length = mean(flipper_length_mm)
  )
```

```
## # A tibble: 3 × 5
##   species   avg_body_mass avg_bill_length avg_bill_depth avg_flipper_length
##   <fct>             <dbl>           <dbl>          <dbl>              <dbl>
## 1 Adelie            3706.            38.8           18.3               190.
## 2 Chinstrap         3733.            48.8           18.4               196.
## 3 Gentoo            5092.            47.6           15.0               217.
```

The results suggest that there are distinct morphological differences among the three species. Gentoo penguins have the highest average body mass and flipper length, while Chinstrap penguins have the longest average bill length. In contrast, Gentoo penguins have the lowest average bill depth, whereas Adelie and Chinstrap penguins have similar values for this trait.

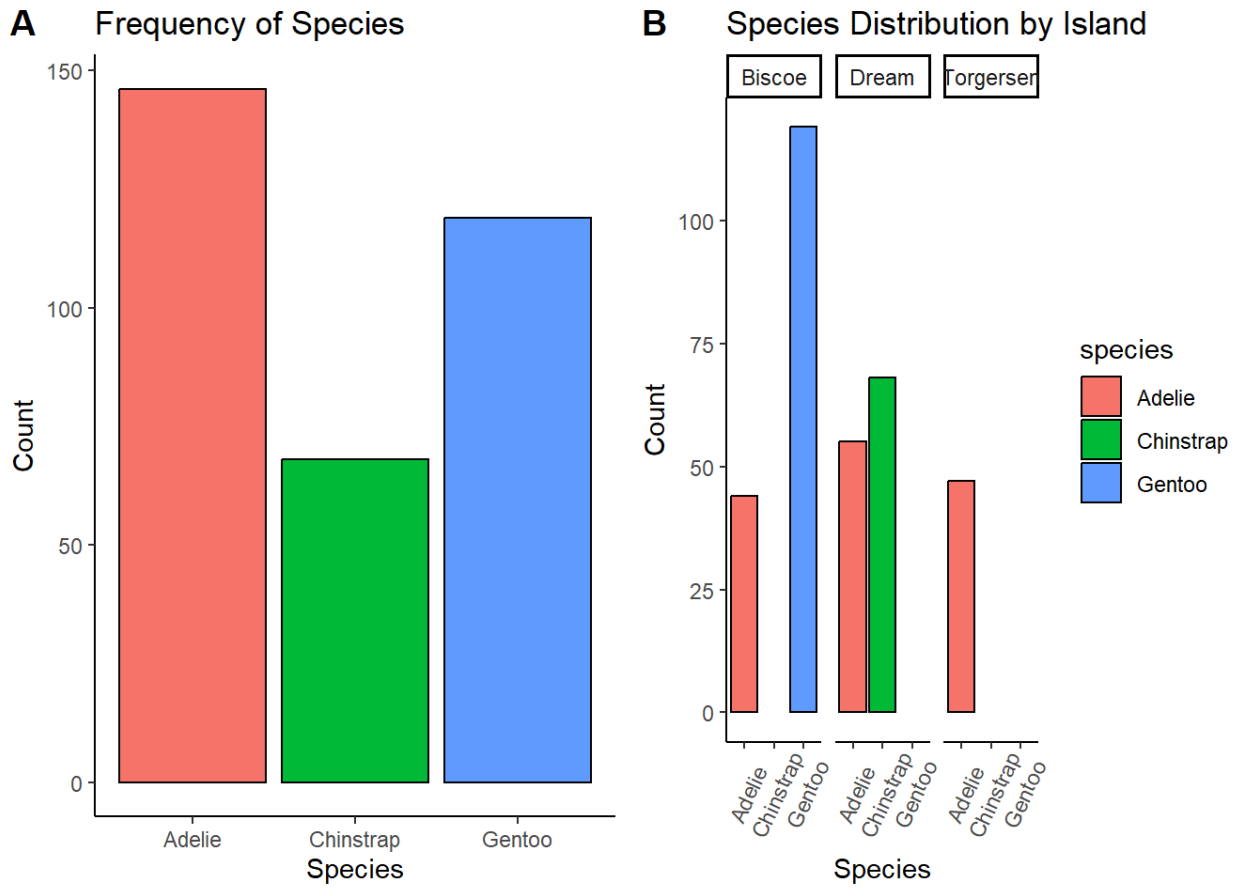
# 4. Exploratory Data Visualisations

## 4.1 Frequency Distribution

The frequency distribution of categorical variables are visualised to further analyse and compare the distribution of data as well as identify any outliers.

### 4.1.1 Species Distribution

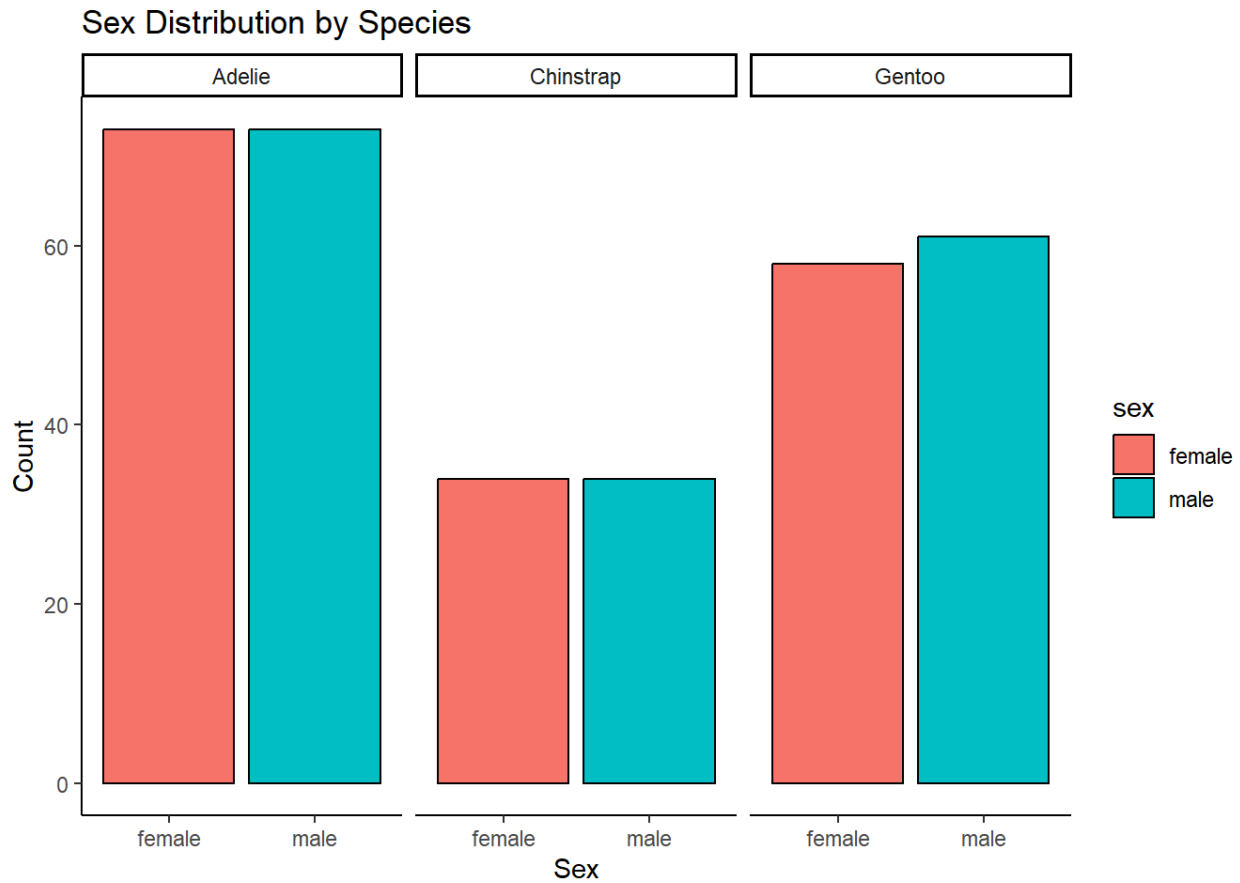
```
# Bar chart showing the count of each species
freqspecies <- ggplot(penguins_clean, aes(x = species, fill = species)) +
  geom_bar(color = "black", show.legend = FALSE) +
  labs(title = "Frequency of Species",
       x = "Species",
       y = "Count")

# Faceted bar chart displaying species distribution across islands
freqspecies_byisland <- ggplot(penguins_clean, aes(x = species, fill = species)) +
  geom_bar(color = "black") +
  facet_wrap( ~ island) +
  labs(title = "Species Distribution by Island",
       x = "Species",
       y = "Count") +
  theme(axis.text.x = element_text(angle=65, vjust=0.6))

# Plotting both bar charts onto a grid
plot_grid(freqspecies, freqspecies_byisland, labels = c("A", "B"), ncol = 2, nrow = 1)
```




As mentioned previously, there is a higher count of Adelie penguins than Gentoo and Chinstrap penguins.

It can be observed that Gentoo penguins are only found on Biscoe island, whereas Chinstrap penguins are only found on Dream island, and Adelie penguins are found on all three islands; Biscoe, Dream, and Torgensen.

### 4.1.2 Sex Distribution by Species

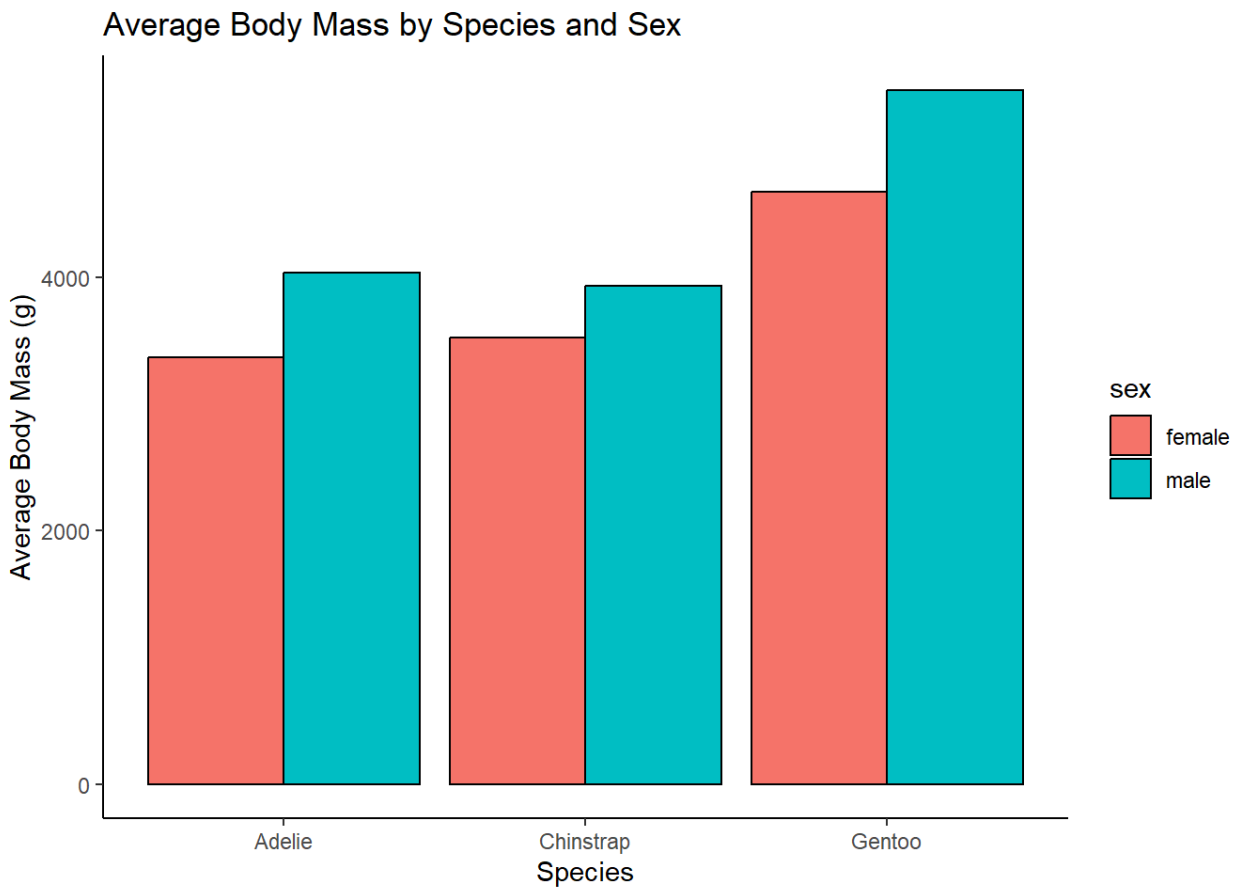


Disregarding the missing values, the distribution of male and female penguins in each species is roughly similar, with a slightly higher count of Gentoo males than Gentoo females.

# 4.2 Body Mass Analysis

## 4.2.1 Average Body Mass by Species and Sex

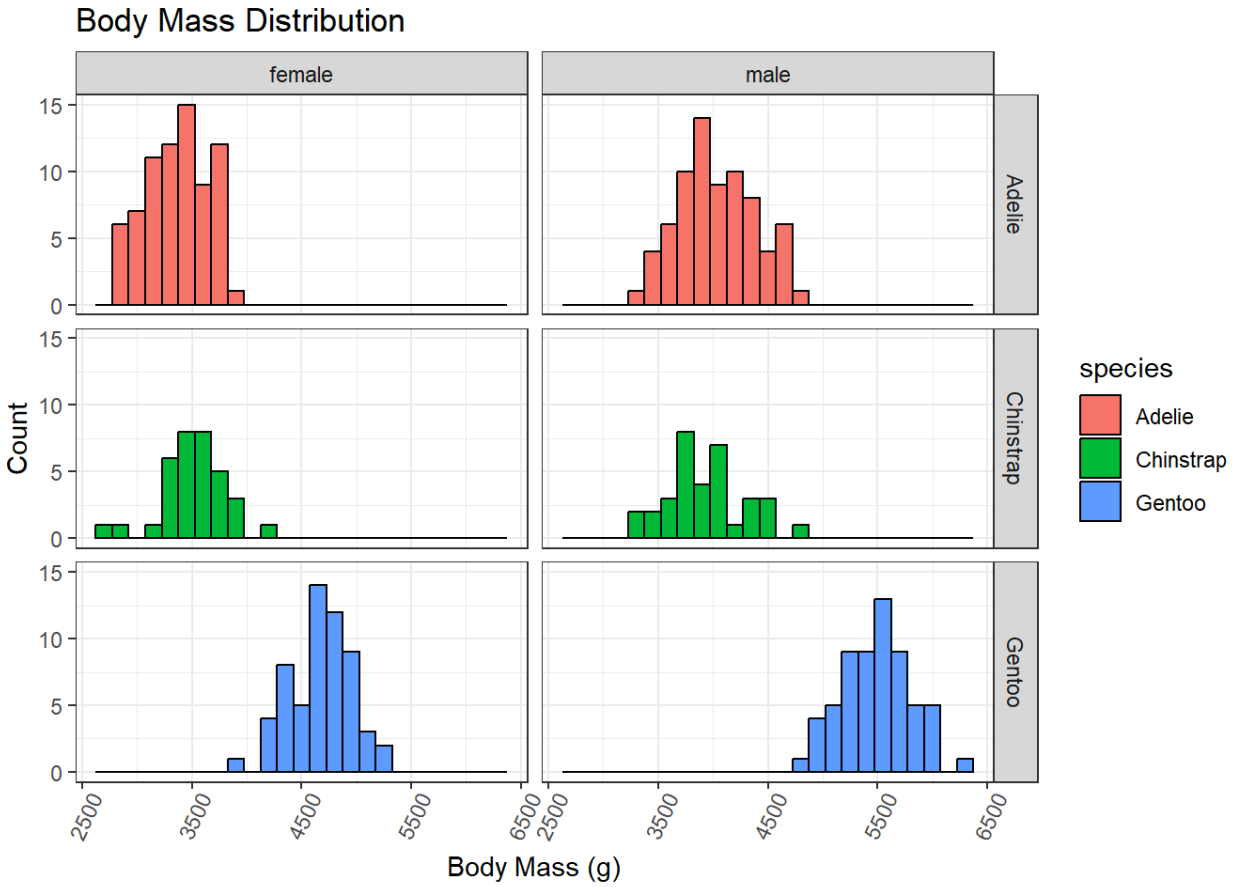
```
# Bar chart showing mean body mass for each species, grouped by sex
ggplot(penguins_clean, aes(x = species, y = body_mass_g, fill = sex)) +
  geom_bar(stat = "summary", fun = "mean", position = "dodge", color = "black") +
  labs(
    title = "Average Body Mass by Species and Sex",
    x = "Species",
    y = "Average Body Mass (g)"
  )
```




Gentoo penguins have the highest average body mass, while Adelie penguins have the lowest. For all three penguin species, males tend to have a higher body mass than females.

## 4.2.2 Histogram of Body Mass Distribution

```
# Histogram of body mass, by species and sex
ggplot(data = penguins_clean, aes(x = body_mass_g, fill = species)) +
  geom_histogram(color = "black", bins = 25) +  # Histogram with black outline
  facet_grid(species ~ sex) +  # Facet by species (rows) and sex (columns) +
  labs(title = "Body Mass Distribution",
       x = "Body Mass (g)",
       y = "Count") +
  theme_bw() +
  theme(axis.text.x = element_text(angle=65, vjust=0.6))
```




***Adelie Penguins***

The distribution of body mass for both females and males appear to be approximately symmetric and unimodal. It is observed that males generally tend to have a slightly higher body mass than females. The peak (which reflects the mode) is well-defined for both sexes.

***Chinstrap Penguins***

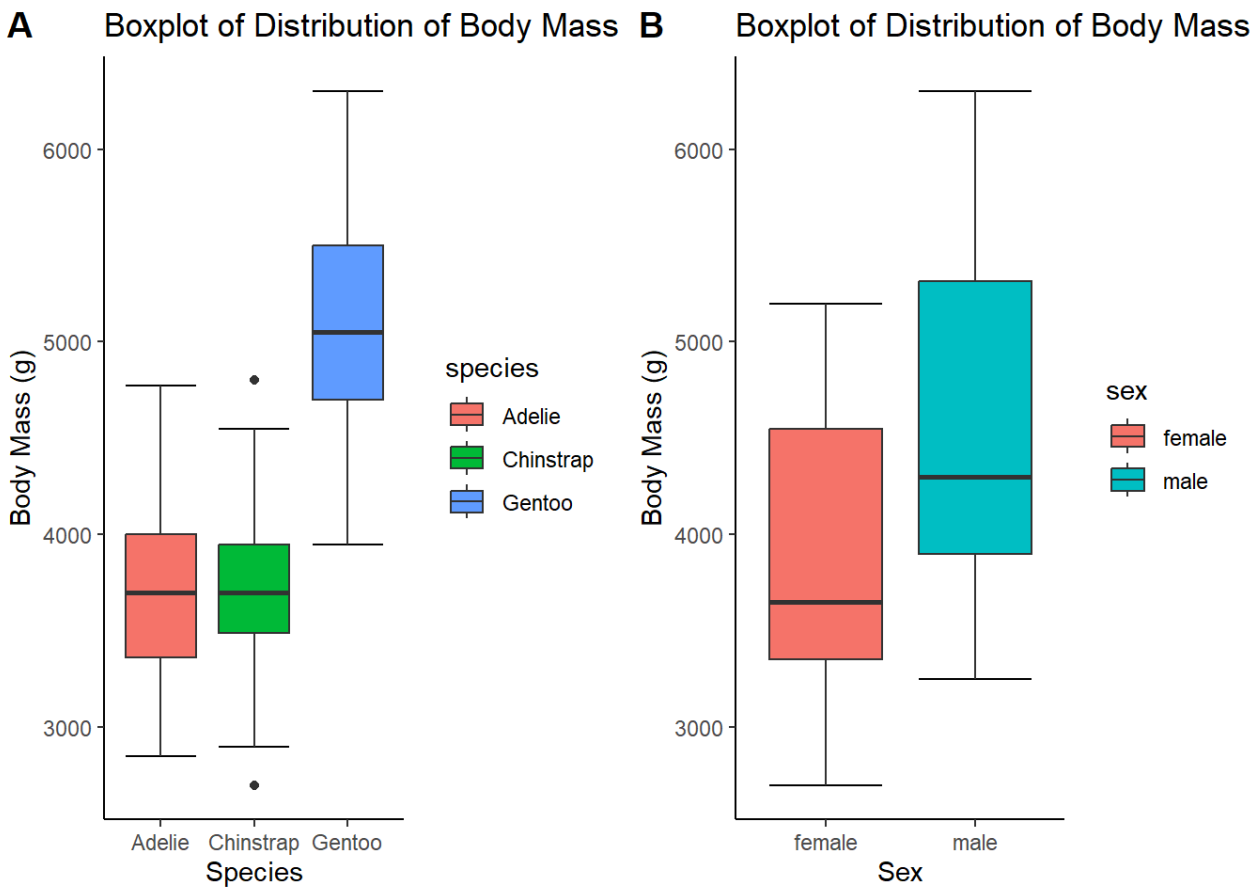
The distribution is less symmetric, showing slight right-skewness particularly for males. The range of the body mass of Chinstrap penguins is concentrated in a narrower range compared to the other two species. Similar to Adelie penguins, the males tend to have a higher body mass than females.

***Gentoo Penguins***

The distribution is approximately normal and unimodal. It is evident that there is a broader range and spread in the data compared to the other two species. Males are significantly heavier than females.

### 4.2.3 Boxplots of Body Mass Distribution

```
# Boxplot showing body mass distribution across species
box_bm_byspec <- ggplot(penguins_clean, aes(x = species, y = body_mass_g, fill = species))
+
  stat_boxplot(geom ='errorbar') +
  geom_boxplot() +
  labs(title = "Boxplot of Distribution of Body Mass (g) by Species",
       x = "Species",
       y = "Body Mass (g)")

# Boxplot showing body mass distribution across sex
box_bm_bysex <- ggplot(penguins_clean, aes(x = sex, y = body_mass_g, fill = sex)) +
  stat_boxplot(geom ='errorbar') +
  geom_boxplot() +
  labs(title = "Boxplot of Distribution of Body Mass (g) by Sex",
       x = "Sex",
       y = "Body Mass (g)")

# Plotting both box plots onto a grid
plot_grid(box_bm_byspec, box_bm_bysex, labels = c("A", "B"), ncol = 2, nrow = 1)
```




***By Species***

From figure A, it is quite evident that the median body mass of Gentoo penguins is the greatest compared to the other two, which values are identical.

Gentoo penguins also have the highest interquartile range, followed by Adelie penguins and Chinstrap penguins. Additionally, Gentoo penguins exhibit the greatest interquartile range, indicating a higher degree of variability in body mass within this species. Two outliers are present in the distribution of body mass of Chinstrap penguins.

***By Sex***

Figure B shows that female penguins tend to have a lower median body mass compared to males. Male penguins have a wider spread of body mass, with a higher median and a larger interquartile range, indicating greater variability in body mass compared to females.

Both data body mass for males and females appear to be right-skewed, as indicated by the longer upper whiskers and the positioning of the median closer to the lower quartile.

# 5. Correlation Analysis

```
##                   Relationship Correlation
## 1    Body Mass and Bill Length   0.5894511
## 2     Body Mass and Bill Depth  -0.4720157
## 3 Body Mass and Flipper Length   0.8729789
```

A strong positive correlation (0.87) is observed between body mass and flipper length, suggesting that as flipper length increases, body mass tends to increase.

A moderate positive correlation (0.59) is seen between body mass and bill length, suggesting that as bill length increases, body mass tends to increase. Bill length may be a secondary predictor of body mass.

A negative correlation (-0.47) exists between body mass and bill depth, suggesting that as bill depth increases, body mass tends to decrease.

Based on these findings, it would be valuable to conduct a scatterplot analysis to examine the relationship between bill length and body mass, as well as flipper length and body mass, since these two variables exhibited the strongest correlations among the three measured traits.

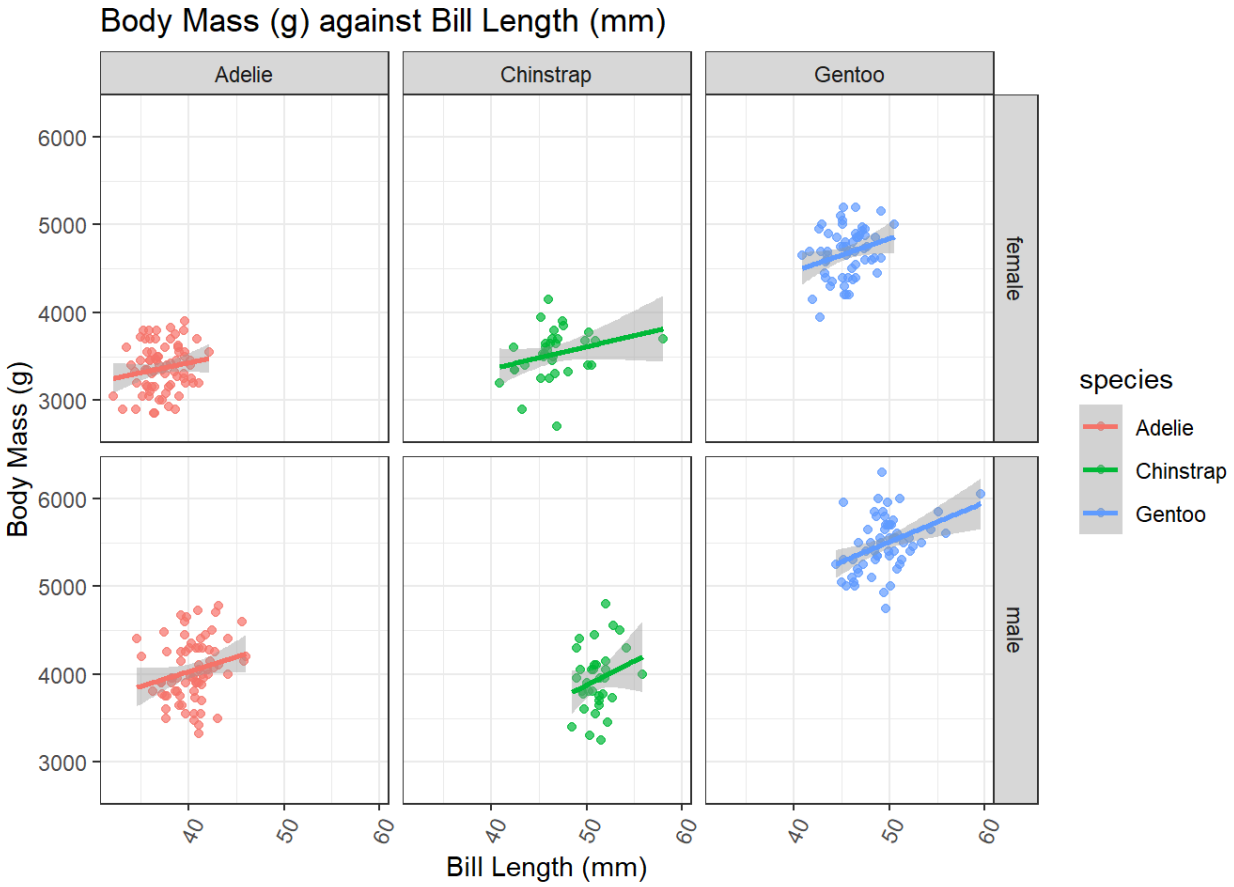
# 6. Scatterplot Analysis

The below scatterplot visualises the relationship between body mass and bill length for the three penguin species, separated by sex.

Across all species, there is a positive correlation evident: as bill length increases, body mass also tends to increase. However, the strength of this relationship is seen to vary among species. The correlation for Adelie penguins in particular appears to be weaker, due to the points being more scattered and the trend lines being less steep. The strongest positive correlation is observed in Gentoo penguins, where the trend line is steeper compared to the other species, suggesting that bill length is a stronger predictor of body mass for this species.

As noted from previous plots, males tend to be heavier than females – although this is most evident in Gentoo penguins.

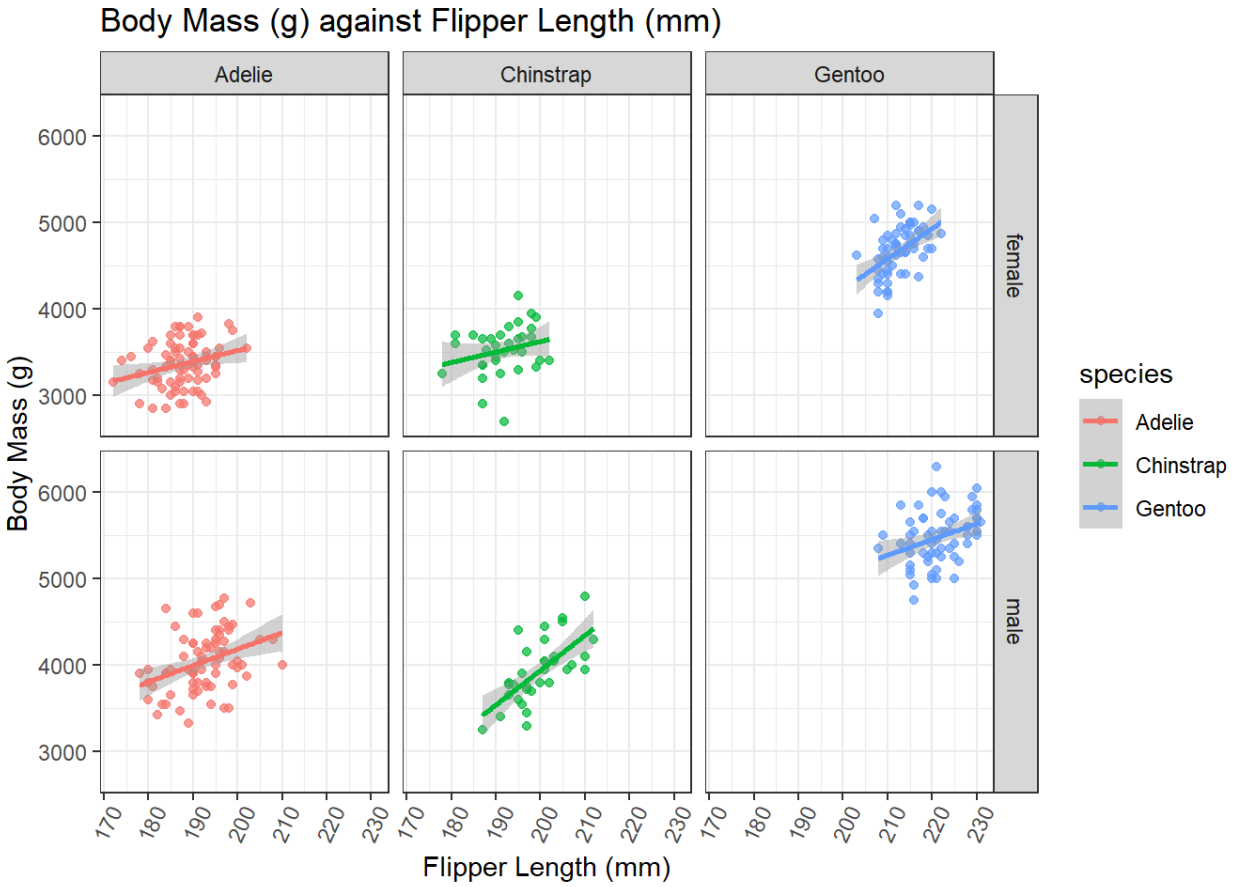
```
# Scatterplots visualizing relationships between Body Mass vs. Flipper Length
# Body Mass vs. Bill Length
ggplot(penguins_clean, aes(x = bill_length_mm, y = body_mass_g, color = species)) +
  geom_point(alpha = 0.7) +
  geom_smooth(method = lm) +
  facet_grid(sex ~ species) +
  labs(title = "Body Mass (g) against Bill Length (mm)",
       x = "Bill Length (mm)",
       y = "Body Mass (g)") +
  theme_bw() +
  theme(axis.text.x = element_text(angle=65, vjust=0.6))
```




The below scatterplot visualises the relationship between body mass and flipper length for the three penguin species, separated by sex.

There is a strong positive correlation between flipper length and body mass across all species. Penguins with longer flippers tend to have higher body mass. As highlighted by the higher correlation, the relationship appears stronger compared to bill length, suggesting that flipper length is a better predictor of body mass.

```
# Body Mass vs. Flipper Length by Species and Sex
ggplot(penguins_clean, aes(x = flipper_length_mm, y = body_mass_g, color = species)) +
  geom_point(alpha = 0.7) +
  geom_smooth(method = lm) +
  facet_grid(sex ~ species) +
  labs(title = "Body Mass (g) against Flipper Length (mm)",
       x = "Flipper Length (mm)",
       y = "Body Mass (g)") +
  theme_bw() +
  theme(axis.text.x = element_text(angle=65, vjust=0.6))
```




# 7. Citation

```
## To cite palmerpenguins in publications use:
##
##   Horst AM, Hill AP, Gorman KB (2020). palmerpenguins: Palmer
##   Archipelago (Antarctica) penguin data. R package version 0.1.0.
##   https://allisonhorst.github.io/palmerpenguins/. doi:
##   10.5281/zenodo.3960218.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {palmerpenguins: Palmer Archipelago (Antarctica) penguin data},
##     author = {Allison Marie Horst and Alison Presmanes Hill and Kristen B Gorman},
##     year = {2020},
##     note = {R package version 0.1.0},
##     doi = {10.5281/zenodo.3960218},
##     url = {https://allisonhorst.github.io/palmerpenguins/},
##   }
```