STAT226: Statistical Computing in R

Predicting Penguin Sex from Morphological Features

Author: Samyak Shrestha

Professor Brian Jones

I) Introduction

Sexual dimorphism in penguins, characterized by differences such as body size and beak dimensions, can influence various biological and ecological aspects such as survival strategies, breeding success, and adaptation to environmental pressures (Fairbairn DJ 1997; Pérez‐Barbería, 2006). Particularly, the Palmer Archipelago of Antarctica presents a unique setting to explore these differences due to it being home to a diverse species of penguins. This project will involve predicting the sex of the penguin using a series of morphometric features such as flipper length, culmen length, culmen depth, and body mass. This difference in the degree and type of sexual dimorphism can provide valuable insights into how physical influence gender characteristics. Furthermore, this project will involve the use of three techniques we have learnt throughout the semester: Bootstrapping, Decision Trees, and Logistic Regression along with specific questions tailored to each.

The research was conducted and data collected were on *Pygoscelis* penguins nesting on several islands within the Palmer Archipelago west of the AP near Anvers Island over the austral summer seasons from 2007 to 2010 by Dr. Kristen Gorman at Palmer Station, which is part of the Antarctica Long Term Ecological Research network (Gorman et al., 2014; Palmer Archipelago Penguin Data, n.d.). The research conducted by Gorman and colleagues aimed to explore the relationship between sex-specific foraging behaviors and changes in the environment. During the breeding season, they collected penguins to conduct body measurements for right flipper length, culmen length, culmen depth, and body mass, and blood tests to determine the sex of the bird.

Table 1 below contains the variables of interest from these data:

Variable	Description
Body mass (g)	This quantitative variable measures the penguin's weight in grams.
Culmen length (mm)	This quantitative variable measures the length of the upper ridge of the penguin's beak.
Culmen depth (mm)	This quantitative variable measures the height of the penguin's bill.
Flipper length (mm)	This quantitative variable measures the length from the penguin's sternum to the tip of its right flipper.
Sex	This binary categorical variable identifies whether the penguin is male (1) or female (0).

Table 1. Variables of interest and their descriptions

II) Exploratory Data Analysis (EDA)

In this section, we explore the Palmer Archipelago penguin dataset to understand the key characteristics and relationships among the variables, particularly focusing on how they differ by sex (Male or Female). By visualizing and summarizing the data, we aim to identify patterns and potential predictors of penguin sex based on their physical measurements.

	sex	mean_culmen_length	mean_culmen_depth	mean_body_mass	mean_flipper_length
	<fct>	<dbl>	<dbl>	<dbl>	<dbl>
1	FEMALE	42.1	16.4	3862.	197.
2	MALE	45.9	17.9	4546.	205.

The summary table above reveals clear differences in physical measurements between male and female penguins. Males have higher mean values across all variables, including culmen length (45.9 mm vs. 42.1 mm), culmen depth (17.9 mm vs. 16.4 mm), body mass (4546 g vs. 3862 g), and flipper length (205 mm vs. 197 mm). These distinctions suggest that body mass, culmen length, and flipper length are the strongest indicators of sex, with culmen depth showing less separation.

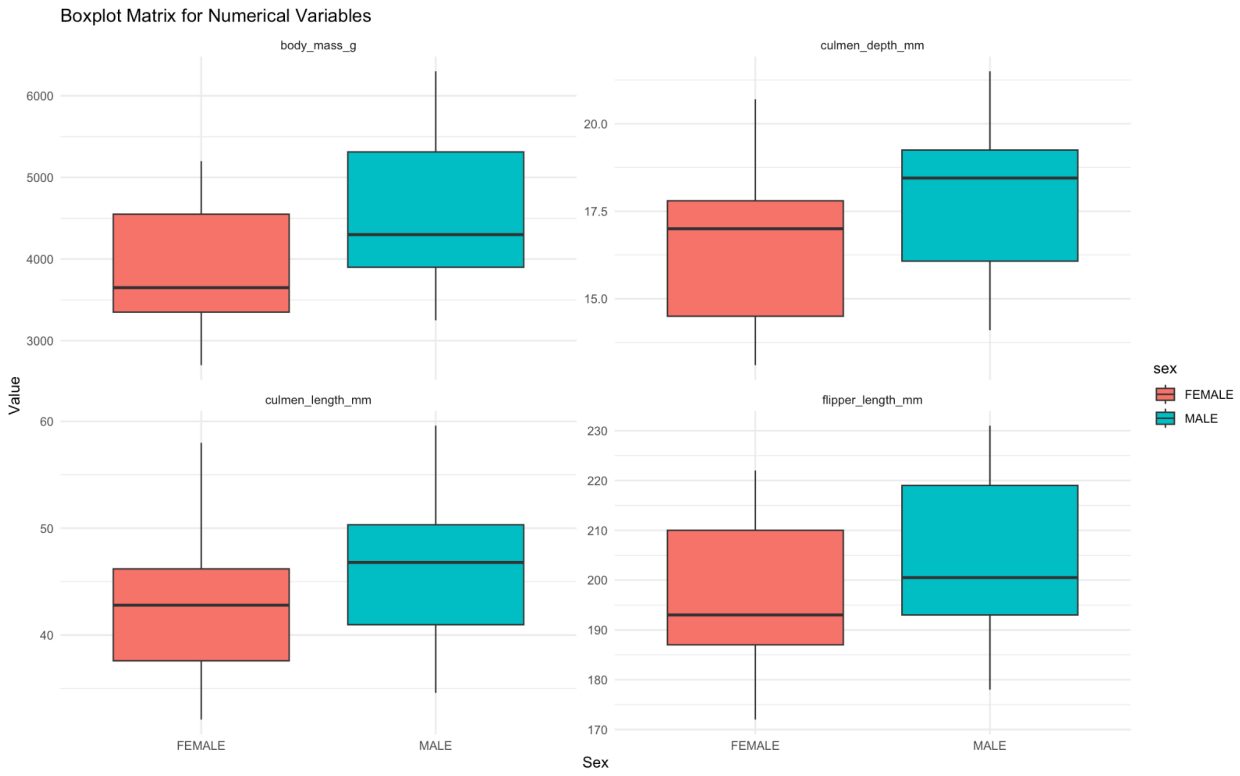

Figure 1. Boxplot matrix of key morphological traits stratified by sex

In Figure 1, we can observe clear differences between males and females across all variables. Males tend to have higher body mass, culmen length, and flipper length, as indicated by their

higher medians and wider distributions compared to females. Culmen depth, however, shows less pronounced separation, with some overlap in values between males and females. These distinctions suggest that body mass, culmen length, and flipper length may serve as strong predictors for determining the sex of penguins, while culmen depth might play a supplementary role.

III) Bootstrapping

1) How does the bootstrap sampling distribution of the mean compare to the original dataset's distribution of body mass? What mathematical concepts are at play?

Body mass is one of the strongest predictors of penguin sex, as demonstrated in the EDA. However, the body mass measurements in the dataset are drawn from a specific sample of penguins. To better understand the variability and reliability of the observed mean body mass, we will construct a bootstrap sampling distribution and compare it to the original dataset.

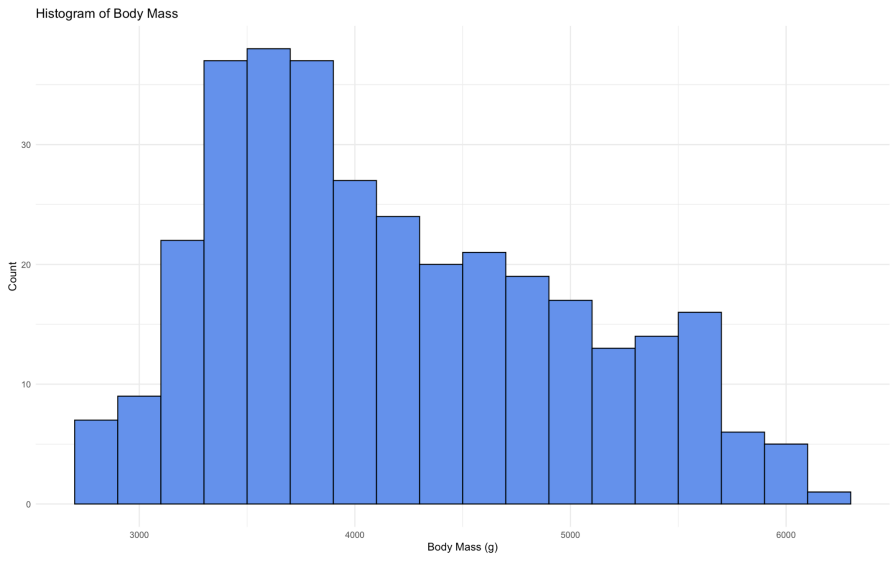

Figure 2. Histogram of Body Mass

Figure 2 depicts the distribution of body mass values for the original dataset. This histogram reflects the actual variation in the penguins' body masses. The data appears slightly skewed to

the right, with most penguins having body mass values concentrated around 4000 to 4500 grams, but with some heavier individuals contributing to a longer tail on the right. To better understand the sampling distribution of the mean, we employ bootstrap resampling, which allows us to approximate the sampling distribution of the mean by repeated sampling with replacement from the observed data.

I used the following R-code to calculate the bootstrap sampling distribution of the mean body mass:

```
num_samp <- 5000
sigma_original <- sd(penguins_samp$body_mass_g)  #standard
deviation of the original sample
resamp_summary <- 1:num_samp %>%
  map_dfr(~ penguins_samp %>%
            slice_sample(n = sampsize, replace = TRUE) %>%
            summarize(
              mean_mass = mean(body_mass_g),
              sd_mass = sd(body_mass_g),
              scale_var = (sampsize - 1) * var(body_mass_g) /
sigma_original^2
            )
```

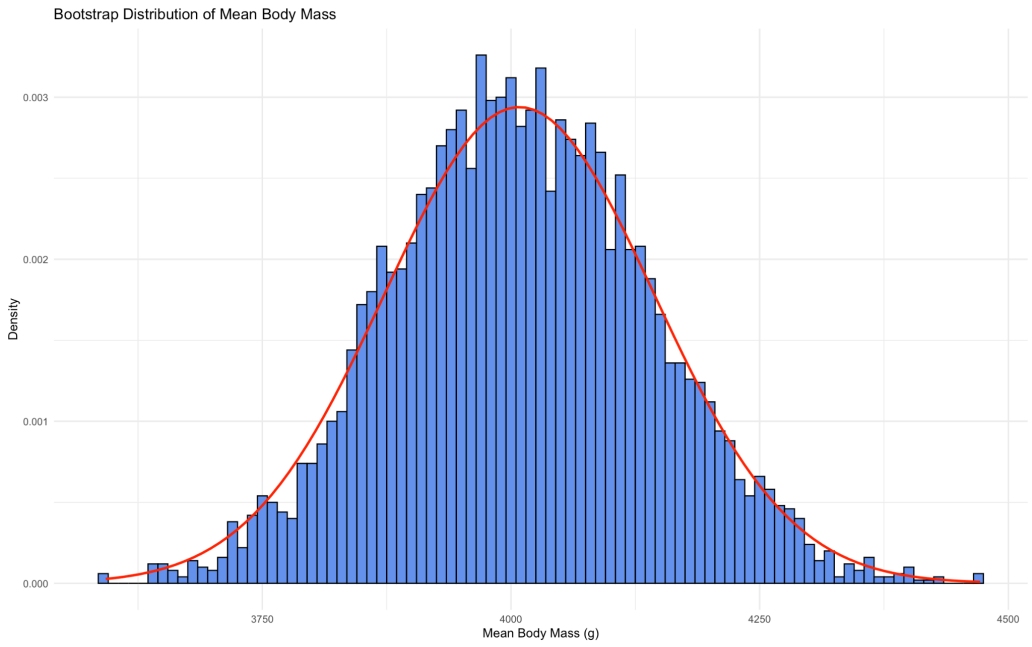

Figure 3. Bootstrap distribution of mean body mass

In Figure 3, each bar corresponds to the mean body mass of one of many bootstrap samples, and the red curve represents a fitted normal distribution. This distribution is much narrower than the original because it reflects the variability of the sample means rather than the variability of individual body masses. There are two mathematical concepts at work here:

i) The **Weak Law of Large Numbers** states that as the sample size (n) increases, the sample mean (X_bar) converges in probability to the population mean (μ):

$$P(|\bar{X} - \mu| > \epsilon) \to 0 \quad \text{as } n \to \infty$$

In Figure 3, the mean of the bootstrap sampling distribution closely approximates the true population mean of body mass because the repeated sampling process simulates large sample sizes. Each bootstrap sample mean clusters around the true mean (μ), with less variability as n increases. This behaviour demonstrates the WLLN, where the mean of the sampling distribution becomes a consistent estimator of the population mean.

ii) The **Central Limit Theorem** states that regardless of the population's original distribution, the sampling distribution of the sample mean (X_bar) approaches a normal distribution as n increases:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

In Figure 3, the bootstrap sampling distribution follows a bell-shaped curve and is approximately normal, even though the original dataset is slightly skewed. This normality arises because each bootstrap sample aggregates many observations, averaging out the population's skewness.

Together, the WLLN and CLT explain why the bootstrap sampling distribution of the mean is centered around the population mean (μ) and has a normal shape. The WLLN ensures that the sample mean converges to the true mean, while the CLT explains why this distribution becomes normal with reduced variance as n increases.

Comparison of favstats

min	Q1	median	Q3	max	mean	sd	n	missing
2700	3550	4050	4775	6300	4207.057	805.2158	333	0

Mean distribution of the original population (up)

min	Q1	median	Q3	max	mean	sd	n	missing
3592.5	3921.667	4007.083	4098.333	4471.667	4010.555	129.9637	5000	0

Bootstrap mean distribution(up)

The original body mass distribution has a mean of 4207.057 g with a standard deviation of 805.2158 g, reflecting the variation in individual penguins' body masses. In contrast, the bootstrap sampling distribution of the mean has a mean of 4010.555 g and a much smaller standard deviation of 129.9637 g, as calculated by resampling the data. This narrower spread reflects the reduced variability in the sample mean compared to individual data points, consistent with the CLT. Additionally, while the original data has a broader range, the bootstrap distribution

of means is more concentrated, highlighting its role in estimating the true population mean with greater precision.

2) Use the bootstrap to generate and interpret a 95% confidence interval for the mean body mass of penguins.

I used the following code to calculate the bootstrap confidence interval:

```
lower1 <- quantile(resamp_summary$mean_mass, 0.025)
upper1 <- quantile(resamp_summary$mean_mass, 0.975)
c(lower1, upper1)
```

```
    2.5%    97.5%
3756.667 4267.521
```

Based on the bootstrapped samples, this interval indicates that we are 95% confident that the true population mean of penguin body mass falls between 3756.667 grams and 4267.521 grams.

IV) Decision Trees

3) What are the most significant physical characteristics (e.g., body mass, culmen length, or culmen depth) for predicting the sex of penguins?

I used the following code to create my decision tree as seen in Figure 4:

```
mod_dtree <- decision_tree(mode = "classification") %>%
  set_engine("rpart") %>%
  fit(sex ~ culmen_length_mm + culmen_depth_mm + body_mass_g +
flipper_length_mm, data = data_tree)
plot(as.party(mod_dtree$fit))
title("Decision Tree for Penguin Sex Prediction")
```

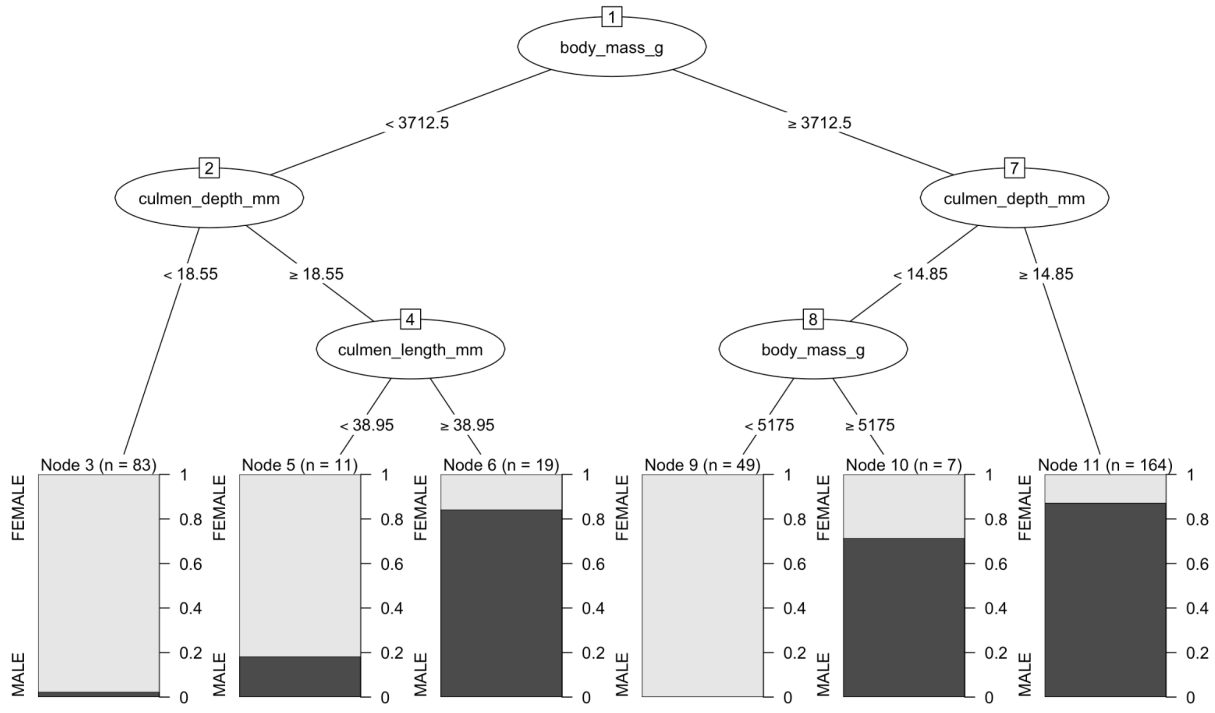

Figure 4. Decision Tree for Penguin Sex Prediction

Based on Figure 4, the decision tree predicts the sex of penguins (Male or Female) based on features like body mass, culmen depth, and culmen length. The most important predictor is body mass, with penguins weighing less than 3712.5 grams generally classified as Female, and those weighing more classified as Male. Culmen depth is the second key feature, where higher values (≥ 18.55 mm) are indicative of Male penguins. Culmen length further refines the classification, with longer culmen lengths (≥ 38.95 mm) more associated with Male penguins. Overall, the tree shows that penguins with lower body mass and smaller culmen dimensions are likely Female, while those with higher body mass and larger culmen dimensions are likely Male.

4) How well does the decision tree classify the sex of penguins, and what is the accuracy of its predictions?

I used the following code to calculate the accuracy and confusion matrix for the decision tree:

```
pred_tree <- data_tree %>%
  bind_cols(
    predict(mod_dtree, new_data = data_tree, type = "class")
  ) %>%
  rename(sex_tree = .pred_class)
accuracy(pred_tree, truth = sex, estimate = sex_tree)
confusion_tree <- pred_tree %>%
  conf_mat(truth = sex, estimate = sex_tree)
print(confusion_tree)
```

.metric	.estimator	.estimate
<chr>	<chr>	<dbl>
accuracy	binary	0.910

	Truth	
Prediction	FEMALE	MALE
FEMALE	139	4
MALE	26	164

The decision tree achieved an accuracy of approximately 0.910 (91%), indicating strong predictive performance. The confusion matrix shows that the model achieves 91% accuracy, correctly classifying 164 males and 139 females, with 30 misclassifications (4 false positives and 26 false negatives).

V) Logistic Regression

Decision trees inherently prioritize variables with the most predictive power, and since flipper length is absent, it likely does not provide substantial predictive information for classifying sex of the penguins. Thus, I choose to drop it from my logistic Regression model. My final model is:

$\text{logit}(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 \text{BodyMass} + \beta_2 \text{CulmenDepth} + \beta_3 \text{CulmenLength}$

5) What is the relationship between culmen depth, culmen length, body mass, and the probability of a penguin being male?

I used the following R code:

```
data_split <- initial_split(penguins, prop = 0.8, strata = sex)
train_data <- training(data_split)
test_data <- testing(data_split)
mod_full <- logistic_reg(mode = "classification") %>%
  set_engine("glm") %>%
  fit(sex ~ culmen_length_mm + culmen_depth_mm + body_mass_g,
data = train_data)
```

```
Call:
stats::glm(formula = sex ~ culmen_length_mm + culmen_depth_mm +
    body_mass_g, family = stats::binomial, data = data)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -5.968e+01  7.807e+00  -7.645 2.10e-14 ***
culmen_length_mm  8.211e-02  4.763e-02   1.724   0.0847 .
culmen_depth_mm   2.027e+00  2.736e-01   7.406 1.30e-13 ***
body_mass_g       5.085e-03  7.223e-04   7.040 1.92e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 368.74  on 265  degrees of freedom
Residual deviance: 128.20  on 262  degrees of freedom
AIC: 136.2

Number of Fisher Scoring iterations: 7
```

Figure 6. Summary output of the Logistic Model

From Figure 6, culmen depth (coefficient = ≅2.027, $p < 0.001$) and body mass (coefficient = 0.005, $p < 0.001$) are highly significant predictors of being male, with increases in these variables strongly associated with higher odds of being male. The coefficient for culmen depth is ≅2.027, which means that for each millimeter increase in culmen depth, the log-odds of being male increase by ≅2.027. Similarly, the coefficient for body mass is ≅0.005, which means that for each one gram increase in body mass, the log-odds of being male increase by ≅0.005. Culmen length has a smaller, marginally significant effect (coefficient = ≅0.082, $p = 0.085$). Again, the coefficient for culmen length is ≅0.082, which means that for each one millimeter increase in culmen length, the log-odds of being male increase by ≅0.082. The model significantly improves over the null model, as indicated by the large reduction in deviance (368.74 to 128.20) and an AIC of 136.2.

6) How accurate is the logistic regression model in predicting the sex of penguins, and how does its performance compare to the decision tree?

I used the following R code:

```
full_accuracy <- accuracy(pred_full, truth = sex, estimate =
sex_full)
confusion_full <- pred_full %>%
  conf_mat(truth = sex, estimate = sex_full)
print(confusion_full)
```

```
# A tibble: 1 × 3
  .metric  .estimator .estimate
  <chr>    <chr>          <dbl>
1 accuracy binary         0.925
```

```
          Truth
Prediction FEMALE MALE
    FEMALE     32    4
    MALE        1   30
```

The logistic regression model achieved an accuracy of approximately 0.925 (92.5%), indicating strong predictive performance on the test dataset. The confusion matrix shows that the model correctly classified 32 females and 30 males, while misclassifying only 5 cases (1 false positive and 4 false negatives).

```
                Model  Accuracy
        Decision Tree 0.9099099
Logistic Regression 0.9253731
```

Overall, both the decision tree and logistic regression models performed well on predicting the sex of the penguin as seen in their accuracy values of over 90%.

VI) Conclusion

In this project, we analyzed the Palmer Archipelago penguin dataset to predict the sex of penguins using physical measurements like body mass, culmen length, flipper length, and

culmen depth. Through exploratory data analysis, we observed distinct differences in these features between males and females. Using decision trees and logistic regression, we achieved strong predictive performance, with accuracies of ≅91% and ≅92.5%. The logistic model indicates that culmen depth is a particularly strong predictor of sex, more so than body mass. This aligns with biological understanding that sexual dimorphism in some bird species can be prominently seen in features like beak depth (Owens, 1988). The exclusion of flipper length from the final model suggests that, although potentially informative, this variable does not provide additional predictive power beyond what body mass and culmen depth already offered in this specific dataset and model setup.Additionally, bootstrap resampling allowed us to estimate the sampling distribution of the mean body mass and construct confidence intervals.

VII) Citations

Gorman, K. B., Williams, T. D., & Fraser, W. R. (2014). Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins (Genus Pygoscelis). *PLOS ONE*, *9*(3), e90081.

https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0090081&type=printable

Pérez‐Barbería, J. (2006). Sexual Segregation in Vertebrates: Ecology of the Two Sexes. Based on a workshop held in Cambridge, September 2002. Edited by K E Ruckstuhl and P Neuhaus. *The Quarterly Review of Biology*, *81*(4), 424–425. https://www.journals.uchicago.edu/doi/10.1086/511617

Fairbairn DJ (1997). Allometry for sexual size dimorphism: pattern and process in the coevolution of body size in males and females. Annu Rev Ecol Syst 28: 659–687. https://www.annualreviews.org/content/journals/10.1146/annurev.ecolsys.28.1.659

Palmer Archipelago (Antarctica) penguin data. (n.d.). from https://www.kaggle.com/datasets/parulpandey/palmer-archipelago-antarctica-penguin-data

Owens, I. P. F., & Hartley, I. R. (1998, March 7). *Sexual dimorphism in birds: Why are there so many different forms of dimorphism?*. Proceedings of the Royal Society B: Biological Sciences. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1688905/

All images were sourced from the open source website “Unsplash” (https://unsplash.com/)