Statistical Analysis of Palmer Penguins: Differentiating Species Using Morphometrics

NAME: M.V. YASHVANTH KUMAR

ROLL NO : 24BCS323

DEPT : CSE – A3

1. Introduction

The Palmer Penguins dataset, first introduced by Dr. Kristen Gorman, has become a modern standard for data science education and statistical analysis. It contains morphometric measurements (physical body measurements) for 344 penguins across three species: Adelie, Chinstrap, and Gentoo.

This paper presents a comprehensive statistical analysis of this dataset. The objective is to move beyond simple data exploration and apply a suite of statistical techniques to model and understand the relationships between the penguins' physical measurements and their categorical traits. We will apply four key techniques:

- Correlation Analysis to understand the relationships between the physical measurements.
- Multiple Linear Regression to model and predict a penguin's body mass.
- Analysis of Variance (ANOVA) to test how factors like species and sex affect body mass.
- Principal Component Analysis (PCA) to reduce the data's dimensionality and visualize how effectively the measurements separate the species.

2. Literature Review

- The methods used in this paper are standard in biology and ecology for "morphometric analysis"—the quantitative study of an organism's physical form.

- On Correlation and Regression: Biologists use regression to understand how one physical trait scales with another (a concept called allometry). For example, studies might use regression to test the relationship between beak shape and body size to understand evolutionary adaptations for feeding.
- On ANOVA: ANOVA is widely used to test for differences in physical traits between groups. For instance, studies use ANOVA to test for "sexual dimorphism," which is a significant morphological difference between males and females of a species.
- On PCA: Principal Component Analysis (PCA) is a cornerstone of morphometric analysis. It is used to take many correlated measurements (like bill length, bill depth, etc.) and reduce them to a few "principal components." This allows researchers to visualize complex shape variations and effectively classify different species or groups in a 2D plot. Our analysis will use these same established techniques.

3. Data Set Collection and Analysis with R Code

3.1 Data Collection

The data was sourced from the palmerpenguins R package, which contains measurements for 344 penguins collected at Palmer Station, Antarctica. The key variables for this analysis are species, island, bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g, and sex.

3.2 Data Cleaning and Analysis with R

The analysis was conducted using R. The palmerpenguins package was loaded, and an initial summary showed 11 missing values for sex and 2 for the other measurements. To ensure the accuracy of the statistical models (which require complete data), all rows with missing values were removed using na.omit(), resulting in a clean dataset of 333 penguins.

3.3 Statistical Techniques (R Code)

```
r

# 1. SETUP: LOAD LIBRARIES AND DATA

# install.packages(c("palmerpenguins", "dplyr", "ggplot2", "corrplot"))
```

```
library(palmerpenguins)
library(dplyr)
library(ggplot2)
library(corrplot)

# Load the dataset
data("penguins")

# 2. DATA CLEANING
penguins_clean <- penguins %>%
  na.omit() # Remove all rows with NA values

# 3. TECHNIQUE 1: CORRELATION ANALYSIS
numeric_penguins <- penguins_clean %>%
  dplyr::select(bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g)
cor_matrix <- cor(numeric_penguins)

# Plot 1: Correlation Plot
corrplot(cor_matrix, method = "number", type = "upper",
         title = "Correlation Matrix of Penguin Measurements",
         mar=c(0,0,1,0))
```

4. TECHNIQUE 2: MULTIPLE LINEAR REGRESSION

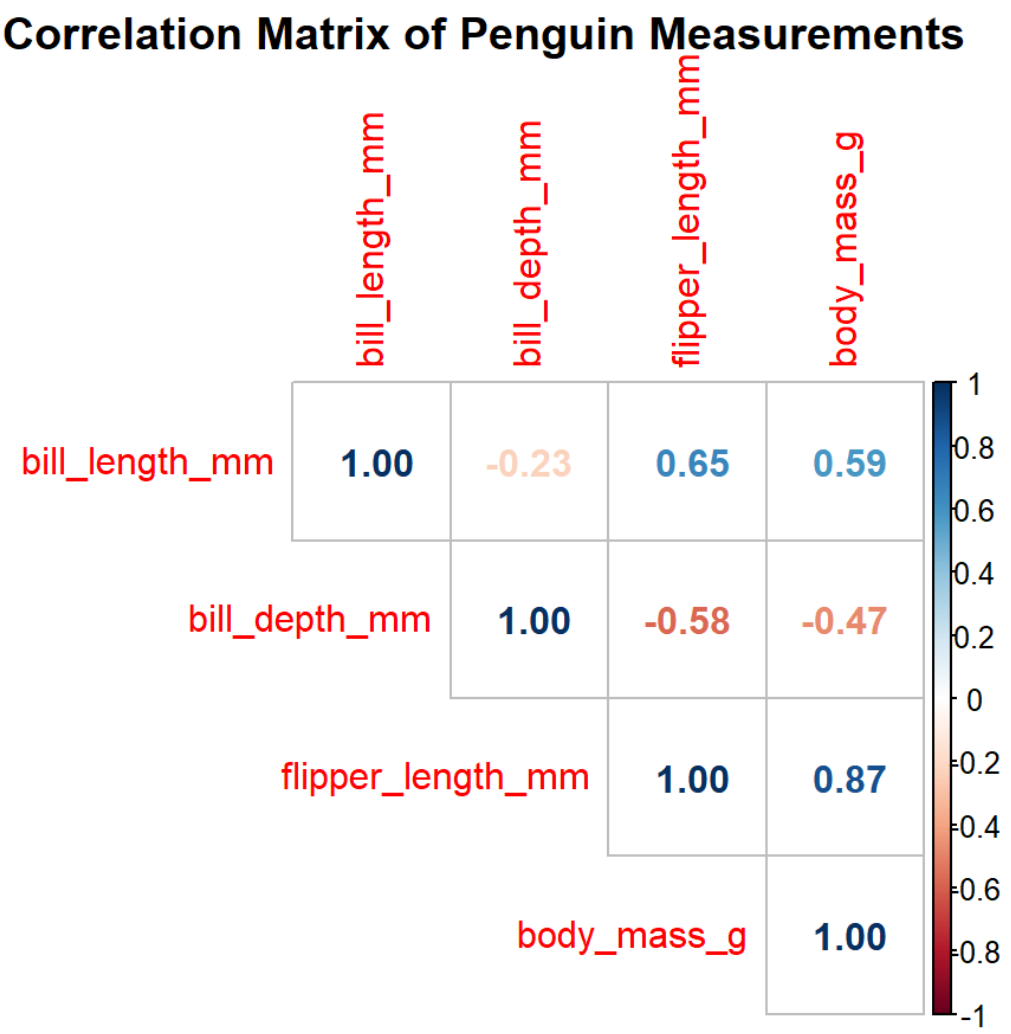
```
# Objective: Predict body_mass_g using flipper_length_mm and species
model <- lm(body_mass_g ~ flipper_length_mm + species, data = penguins_clean)
print(summary(model))

# Plot 2: Regression Scatter Plot
plot_regr_mass <- ggplot(penguins_clean, aes(x = flipper_length_mm, y = body_mass_g,
color = species)) +
```

```
 geom_point(alpha = 0.6) +
 geom_smooth(method = "lm", se = FALSE) +
 labs(title = "Body Mass as a Function of Flipper Length and Species",
      x = "Flipper Length (mm)", y = "Body Mass (g)") +
 theme_minimal()
print(plot_regr_mass)

# 5. TECHNIQUE 3: ANALYSIS OF VARIANCE (ANOVA)
# Objective: Test if body_mass_g differs by species, sex, and their interaction
anova_model <- aov(body_mass_g ~ species * sex, data = penguins_clean)
print(summary(anova_model))
print(TukeyHSD(anova_model))

# Plot 3: Box Plot for ANOVA
plot_anova_box <- ggplot(penguins_clean, aes(x = species, y = body_mass_g, fill = sex)) +
 geom_boxplot() +
 labs(title = "Body Mass by Species and Sex",
      x = "Species", y = "Body Mass (g)") +
 theme_minimal()
print(plot_anova_box)

# 6. TECHNIQUE 4: PRINCIPAL COMPONENT ANALYSIS (PCA)
pca_result <- prcomp(numeric_penguins, scale. = TRUE)
print(summary(pca_result))

# Plot 4: PCA Plot
pca_data <- as.data.frame(pca_result$x)
pca_data$species <- penguins_clean$species
plot_pca <- ggplot(pca_data, aes(x = PC1, y = PC2, color = species)) +
```

```
  geom_point(size = 3, alpha = 0.8) +
  labs(title = "PCA of Penguin Measurements",
       x = paste("PC1 (", round(summary(pca_result)$importance[2,1]*100, 1), "%)", sep=""),
       y = paste("PC2 (", round(summary(pca_result)$importance[2,2]*100, 1), "%)", sep="")) +
  theme_minimal()
print(plot_pca)
```

4. Results

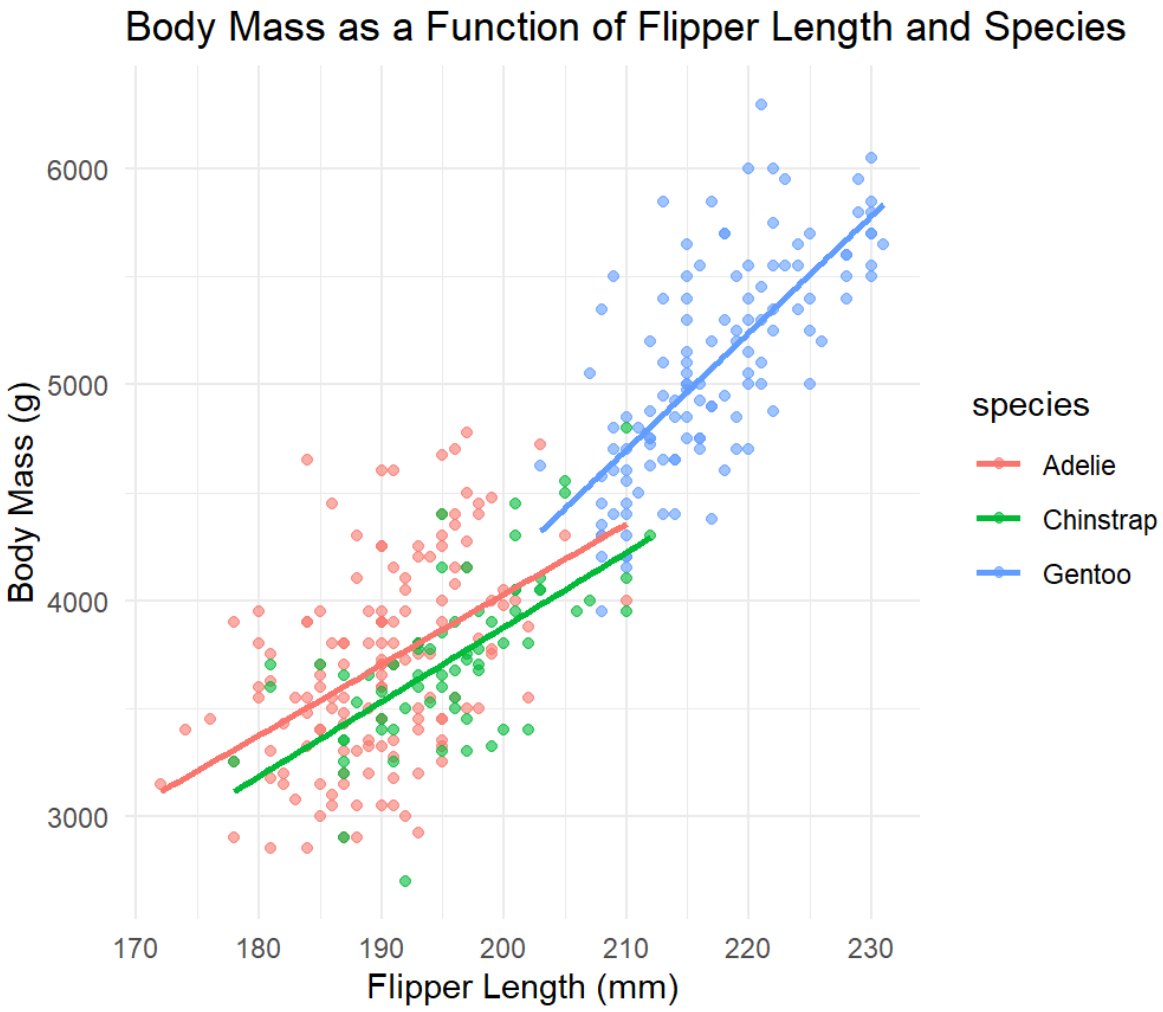
The R output provided clear results for all four analyses.



Technique 1: Correlation Analysis

- The correlation matrix (Plot 1) revealed strong relationships between the physical measurements:

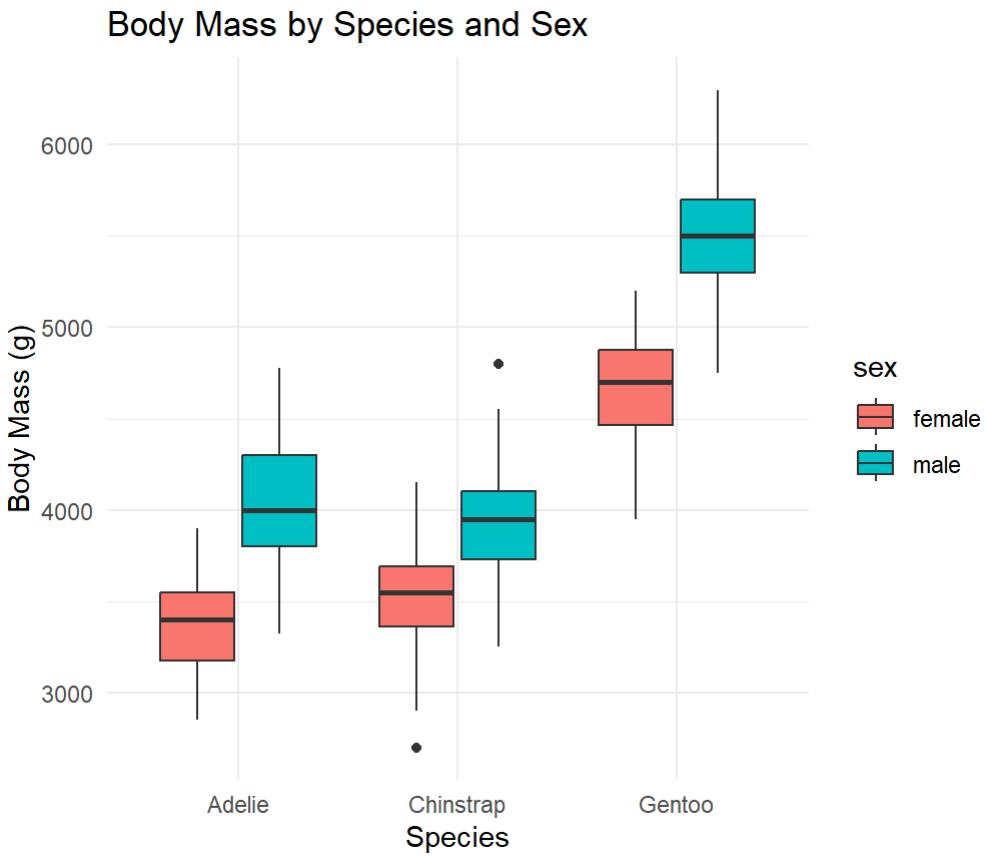
- body_mass_g and flipper_length_mm are very strongly and positively correlated (0.873). This makes sense: penguins with longer flippers are generally larger.
- bill_length_mm and flipper_length_mm** are also strongly correlated (0.653).
- bill_depth_mm is negatively correlated with flipper_length_mm (-0.578) and body_mass_g (-0.472**), suggesting penguins with deeper bills tend to be smaller overall.



Technique 2: Multiple Linear Regression

- The regression model built to predict body_mass_g was highly successful.

- Model Fit: The Adjusted R-squared was *0.7851, meaning that **78.5% of the variance in penguin body mass* can be explained by just its flipper length and species. The overall model was extremely significant (p-value \< 2.2e-16).
- Key Predictors:
- flipper_length_mm (p \< 2e-16): This is a powerful predictor. For every 1mm increase in flipper length, the penguin's body mass is predicted to increase by *40.61 grams*.
- speciesGentoo (p = 0.003): Gentoos are significantly different. On average, a Gentoo penguin is *284.52 grams heavier* than an Adelie penguin (the baseline), even after accounting for their flipper length.
- speciesChinstrap (p = 0.0004): Chinstraps are also different, being *205.38 grams lighter* than an Adelie penguin on average, holding flipper length constant.

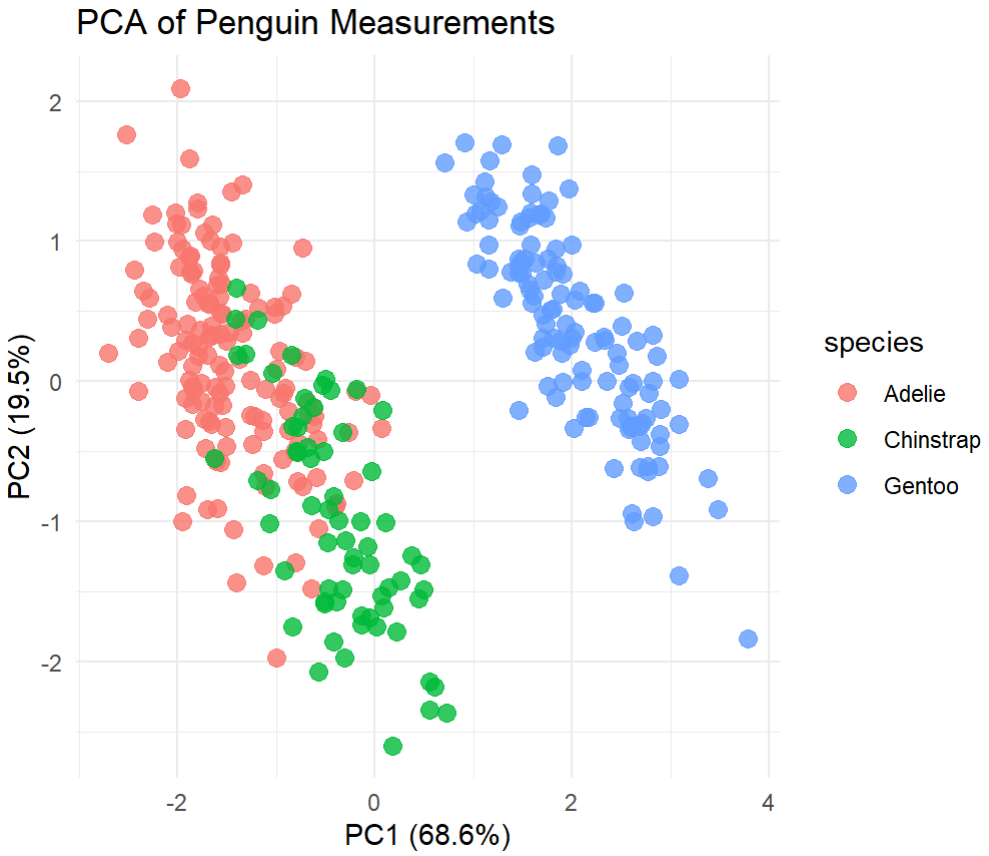


Technique 3: Two-Way ANOVA

The ANOVA test on body_mass_g found that all factors were significant:

- species (p \< 2e-16): A penguin's *species* has a highly significant effect on its body mass.
- sex (p \< 2e-16): A penguin's ‘sex ‘ also has a highly significant effect on its body mass. The Tukey test showed males are, on average, 667g heavier than females.

- species:sex Interaction (p = 0.000197): This is a key finding. The p-value is highly significant, meaning there is a strong *interaction effect. This tells us that the effect of sex on body mass is *not the same for all three species. As seen in Plot 3, the mass difference between males and females is much larger for Gentoo penguins than it is for Adelie or Chinstrap penguins.



Technique 4: Principal Component Analysis (PCA)

- The PCA successfully reduced the four physical measurements into components that explain species differences.

- Variance Explained: The PCA Summary shows that the first two components (PC1 and PC2) combined explain 88.09% of the total variance in the dataset.
- Visualization: The PCA Plot (Plot 4) is the most powerful visualization. It shows that these two components effectively *separate the three species into distinct, non-overlapping clusters.
- The Gentoo penguins (green) are clearly separated from the others on the left side (low PC1 values).
- The Adelie (pink) and Chinstrap (blue) penguins are separated from each other along the PC2 axis. This plot proves that the physical measurements, when combined, are highly effective at differentiating the three species.

5. Reference

- Close, R. A., & Rayfield, E. J. (2012). Functional Morphometric Analysis of the Furcula in Mesozoic Birds. PLOS ONE.
- 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. (This is the original paper for the dataset)
- Horst, A., Hill, A., & Gorman, K. (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data (R package version 0.1.1). https://allisonhorst.github.io/palmerpenguins/
- Horst, A., Hill, A., & Gorman, K. (2022). Palmer Archipelago Penguins Data in the palmerpenguins R Package: An Alternative to Anderson's Irises. The R Journal.
- Obo, U. F., & Bilge, G. (2025). Geometric Morphometric Analysis of Sexual Dimorphism in the Bill of the White Stork. MDPI.
- ResearchGate. (2016). Bird Species Identification System Using Kernel based PCA.
- Wang, X. et al. (2021). Quantitative Analysis of Morphometric Data of Pre-modern Birds: Phylogenetic Versus Ecological Signal. Frontiers in Earth Science.
- Yu, H. et al. (2021). Evolution of body morphology and beak shape revealed by a morphometric analysis. ResearchGate.