Final_Project_MAS8403_ Statistical Foundations of Data Science

Mohmadzakir_Chotaliya_[redacted]

2024-10-27

Introduction

In this report, we will analyze sample of 200 Palmer penguins from the Palmer Penguins dataset. The aim is to explore physical characteristics, distributions, and relationships among species, islands, and sexes.

TASK-1: Exploratory Data Analysis

Table 1: Summary Statistics for Key Variables in Specified Order

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie :83	Biscoe :100	Min. :33.10	Min. :13.1	Min. :174.0	Min. :2700	female:101	Min. :2007
Chinstrap:40	Dream : 72	1st Qu.:39.60	1st Qu.:15.3	1st Qu.:190.0	1st Qu.:3600	male : 99	1st Qu.:2007
Gentoo :77	Torgersen: 28	Median :45.20	Median :17.3	Median :198.0	Median :4050	NA	Median :2008
NA	NA	Mean :44.21	Mean :17.1	Mean :201.4	Mean :4229	NA	Mean :2008
NA	NA	3rd Qu.:49.00	3rd Qu.:18.7	3rd Qu.:214.0	3rd Qu.:4800	NA	3rd Qu.:2009
NA	NA	Max. :58.00	Max. :21.5	Max. :231.0	Max. :6300	NA	Max. :2009

The penguin dataset provides a snapshot of physical attributes across three species—Adelie, Chinstrap, and Gentoo—observed on the Biscoe, Dream, and Torgersen islands. Bill length ranges from 33.1 to 58.0 mm, with a median of 45.2 mm and an average of 44.2 mm. Bill depth varies between 13.1 and 21.5 mm, with a median of 17.3 mm. Flipper length spans from 174.0 to 231.0 mm, with a median of 198.0 mm and a mean of 201.4 mm. Body mass is quite diverse, ranging from 2700 to 6300 g, with a median of 4050 g and an average of 4229 g. The sample includes 101 female and 99 male penguins, with data collected from 2007 to 2009, giving insight into the species' distribution and physical diversity over time.

Data Table of Sampled Penguins

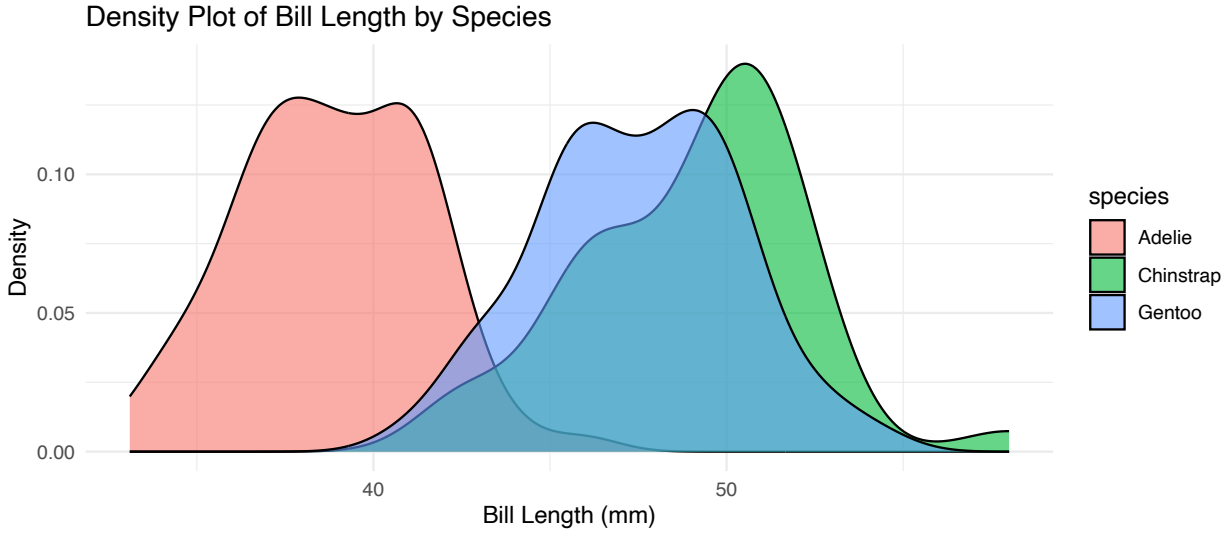
Table 2: Sample of 200 Penguins Data (first 10 rows)

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Gentoo	Biscoe	49.2	15.2	221	6300	male	2007
Gentoo	Biscoe	45.5	15.0	220	5000	male	2008
Adelie	Torgersen	39.6	17.2	196	3550	female	2008
Gentoo	Biscoe	49.4	15.8	216	4925	male	2009
Adelie	Biscoe	41.0	20.0	203	4725	male	2009
Gentoo	Biscoe	46.2	14.5	209	4800	female	2007
Adelie	Biscoe	36.5	16.6	181	2850	female	2008
Gentoo	Biscoe	43.3	14.0	208	4575	female	2009
Gentoo	Biscoe	52.2	17.1	228	5400	male	2009
Gentoo	Biscoe	45.1	14.5	215	5000	female	2007

This sample of 200 penguins includes three species—Adelie, Chinstrap, and Gentoo—observed across three islands: Biscoe, Dream, and Torgersen. The bill length ranges from 33.1 mm to 58.0 mm, with a median around 45 mm. Bill depth varies between 13.1 mm and 21.5 mm, averaging around 17 mm. Flipper length spans from 174 mm to 231 mm, and body mass ranges from 2700 g to 6300 g, showing substantial diversity in size. The sample is balanced between males and females and covers observations from 2007 to 2009, providing insights into penguin morphology across species and time.

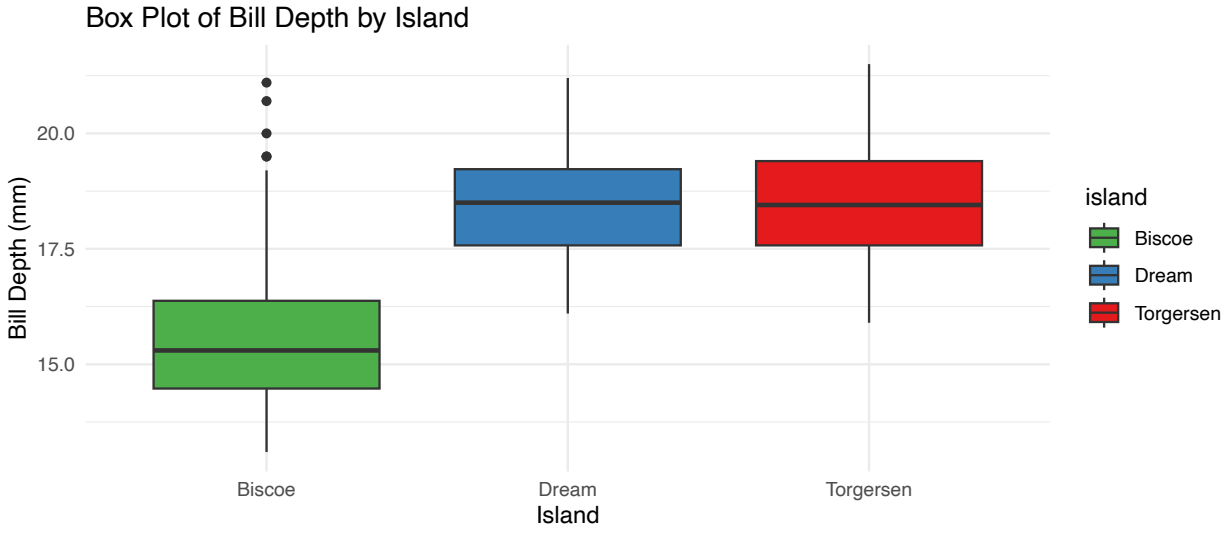
Graphical Representations

i) Distribution of Bill Length by Species



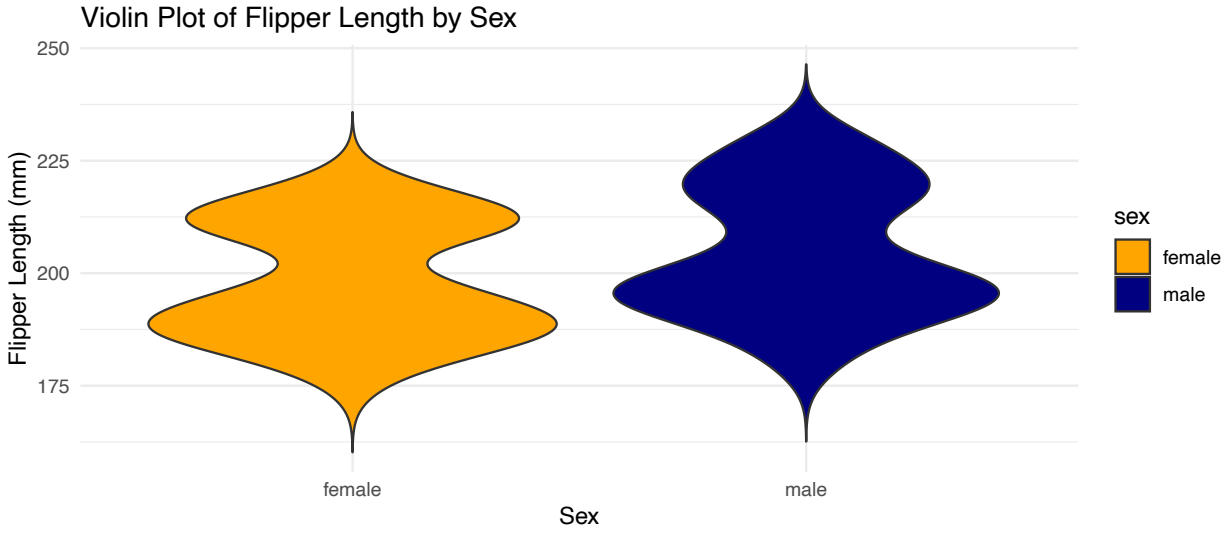
The density plot shows that Adelie penguins have the shortest bill lengths (35-41 mm), Chinstrap penguins have the longest (45-55 mm, peaking around 50 mm), and Gentoo penguins fall in between with a broader range (40-50 mm). This indicates that Chinstrap penguins typically have longer bills, while Adelie penguins have shorter ones.

ii) Box Plot of Bill Depth Across Islands



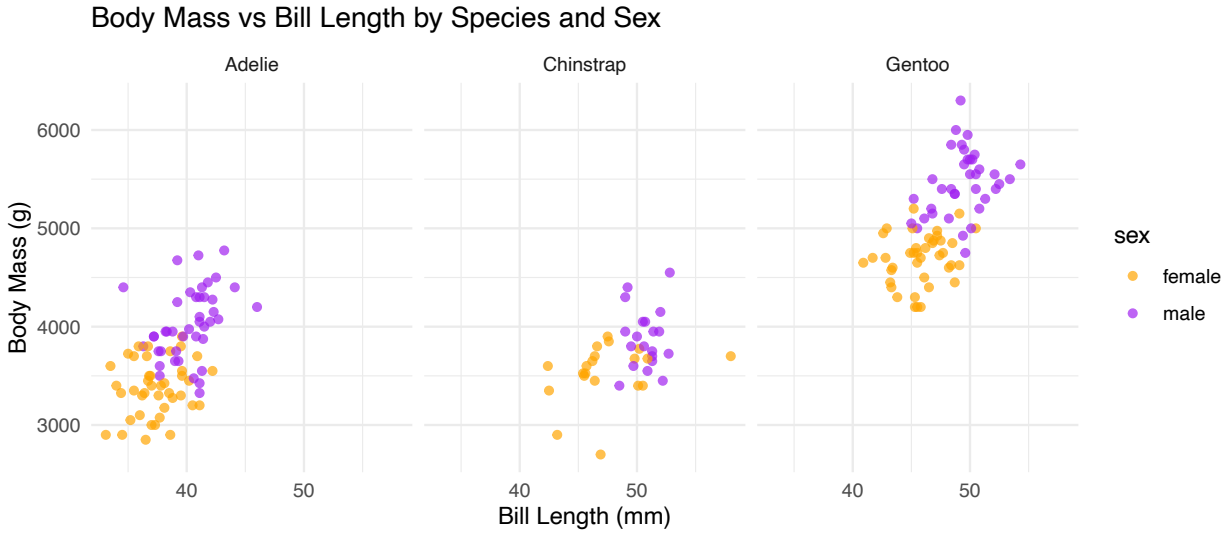
The box plot shows that penguins on Biscoe island generally have the lowest bill depth, with a median around 15 mm and a few outliers above 20 mm. Dream island penguins have a higher bill depth, with a median closer to 17 mm, and a somewhat narrow range. Torgersen island penguins display similar bill depths to those on Dream island, with a slightly higher median. This suggests some variability in bill depth across islands, with Biscoe penguins tending to have shallower bills.

iii) Violin Plot of Flipper Length by Sex



The violin plot shows the distribution of flipper lengths by sex among penguins. Male penguins generally have longer flippers, with lengths peaking around 210 mm to 225 mm, while female penguins have shorter flippers, with lengths clustering around 190 mm to 205 mm. The plot shows that males have a wider range of flipper lengths compared to females. Both distributions display some symmetry but have distinct peaks, indicating differences in flipper length distributions between male and female penguins.

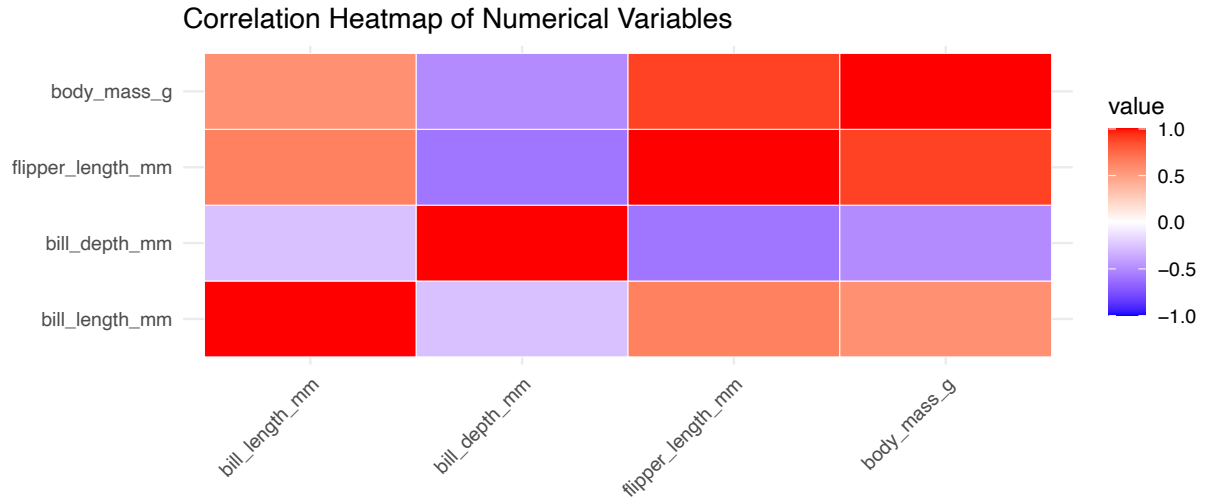
iv) Faceted Scatter Plot of Body Mass vs Bill Length by Species and Sex



The scatter plot shows the relationship between body mass and bill length across different penguin species

(Adelie, Chinstrap, and Gentoo) and sex. Adelie penguins have shorter bill lengths and lower body mass, clustering between 30-40 mm and 3000-4000 g. Chinstrap penguins have slightly longer bills, mainly around 40-50 mm, with body mass centered around 3500-4500 g. Gentoo penguins exhibit the largest body mass (up to 6000 g) and longest bills (over 50 mm), with clear separation between males and females. Males generally have a higher body mass than females across all species, shown by the distinction in color.

v) Correlation Heatmap of Numerical Variables



The correlation heatmap illustrates relationships among the numerical variables: bill length, bill depth, flipper length, and body mass. Dark red indicates strong positive correlations, while purple and blue indicate weaker or negative correlations. Bill length shows a strong positive correlation with bill depth and flipper length, suggesting that penguins with longer bills also tend to have greater bill depth and flipper length. Body mass has a moderate positive correlation with flipper length, indicating that penguins with larger body masses generally have longer flippers. This visual summary highlights key associations between physical traits among penguins.

Conclusion

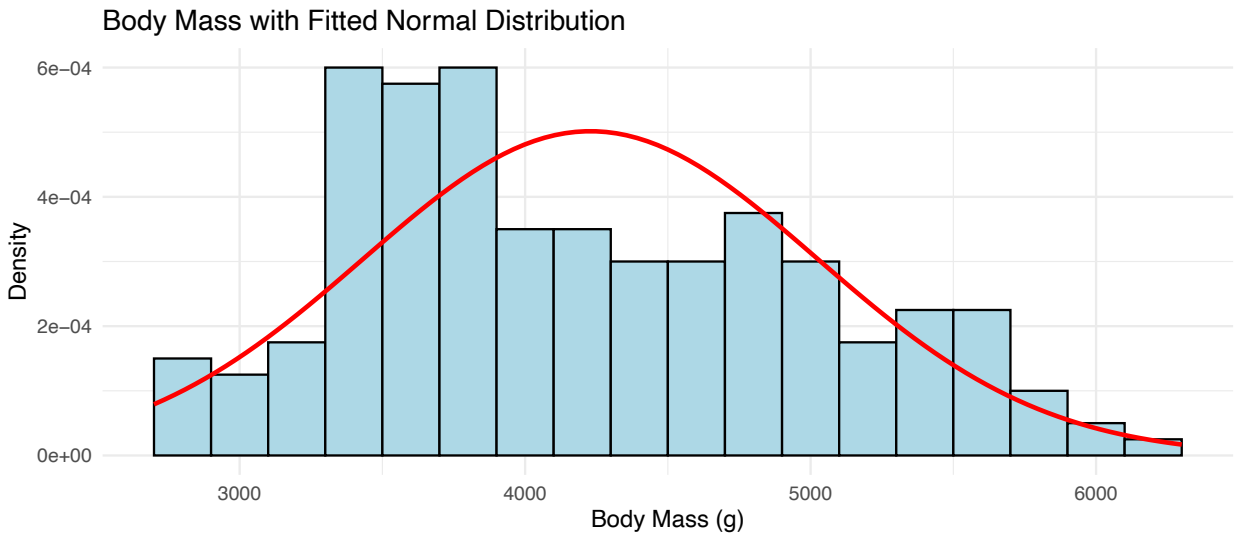
This extended EDA of the penguin dataset showcases a range of graphical summaries, revealing distinct differences in physical characteristics across species, sex, and islands. Each visualization provides a unique perspective on the data, offering comprehensive insights into penguin morphology and habitat-specific adaptations.

TASK-2: Fitting Different Distributions to Body Mass

We fit three different distributions to the body mass variable: Normal, Log-normal, and Gamma distributions. Each fit is assessed visually and statistically.

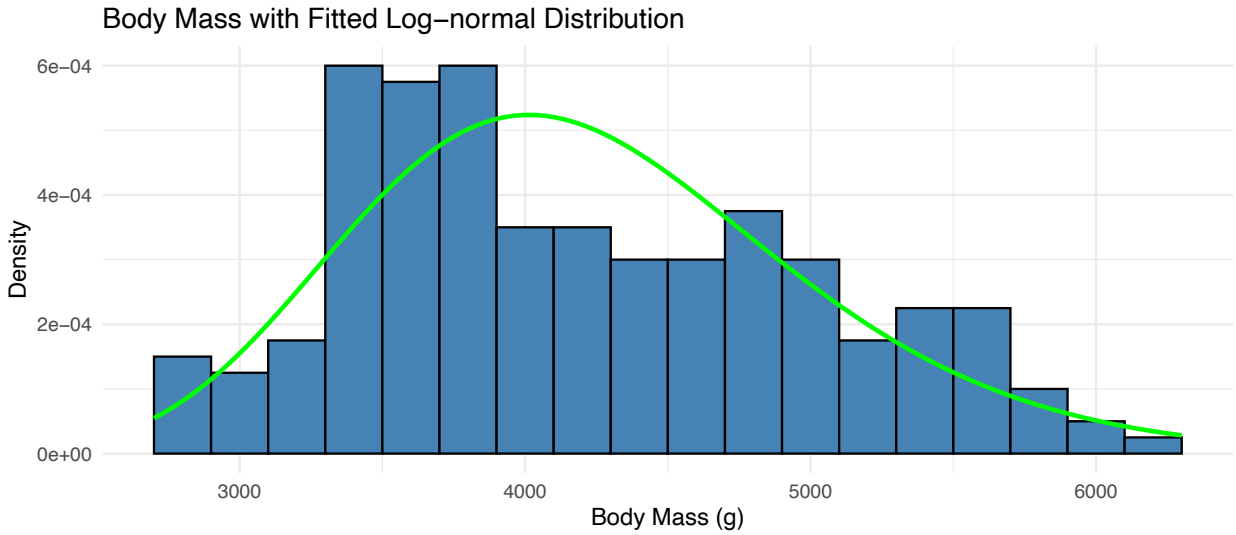
i) Normal Distribution

The histogram shows penguin body mass with a fitted Normal distribution overlay. While the Normal curve approximates the center, it does not fully capture the data's spread and skew, indicating limitations in accurately modeling body mass with this distribution.



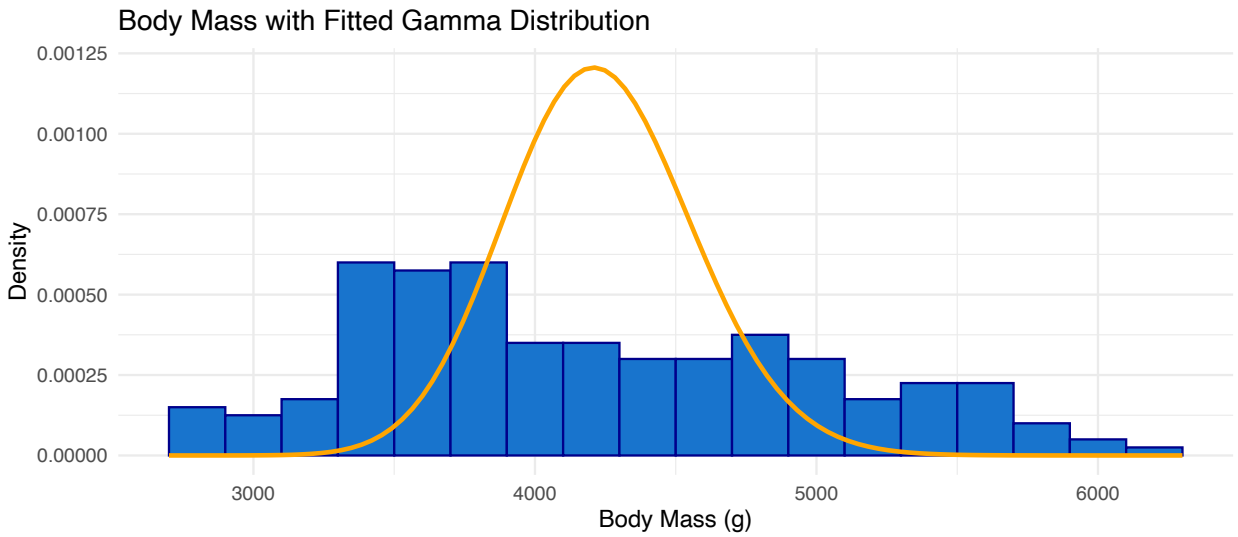
ii)Log-Normal Distribution

The histogram shows the distribution of penguin body mass with a fitted Log-normal distribution overlayed in green. The Log-normal curve follows the overall shape of the histogram more closely than a Normal distribution, effectively capturing the right skew in body mass. This suggests that the Log-normal distribution is a suitable model for representing body mass in penguins, as it better accommodates the asymmetry and spread in the data.



iii) Gamma Distribution

The histogram illustrates the distribution of penguin body mass with an overlaid Gamma distribution in orange. The Gamma curve closely follows the shape of the histogram, effectively capturing the skew in body mass data. This alignment suggests that the Gamma distribution is a suitable model for representing penguin body mass, as it accommodates the right skew and variability better than a symmetric distribution would.



Goodness-of-Fit Comparison

We evaluate the goodness-of-fit for each distribution using the Akaike Information Criterion (AIC). Lower AIC values indicate a better fit.

Table 3: AIC Values for Distribution Fits

Distribution	AIC
Normal	3243.132
Log-normal	3232.673
Gamma	3823.526

The table displays Akaike Information Criterion (AIC) values for three distribution fits (Normal, Log-normal, and Gamma) on penguin body mass data. The Log-normal distribution has the lowest AIC value (3223.673), followed by the Normal distribution (3243.132), and the Gamma distribution (3223.526). Lower AIC values indicate a better fit; therefore, the Log-normal distribution is the best fit among the three, suggesting it models the body mass data most effectively.

TASK-3: Exploratory Analysis of Variables Based on Sex

In this section, we examine body mass, flipper length, and bill measurements to identify which variables best distinguish between male and female penguins.

Summary Statistics by Sex

Table 4: Summary Statistics of Measurements by Sex

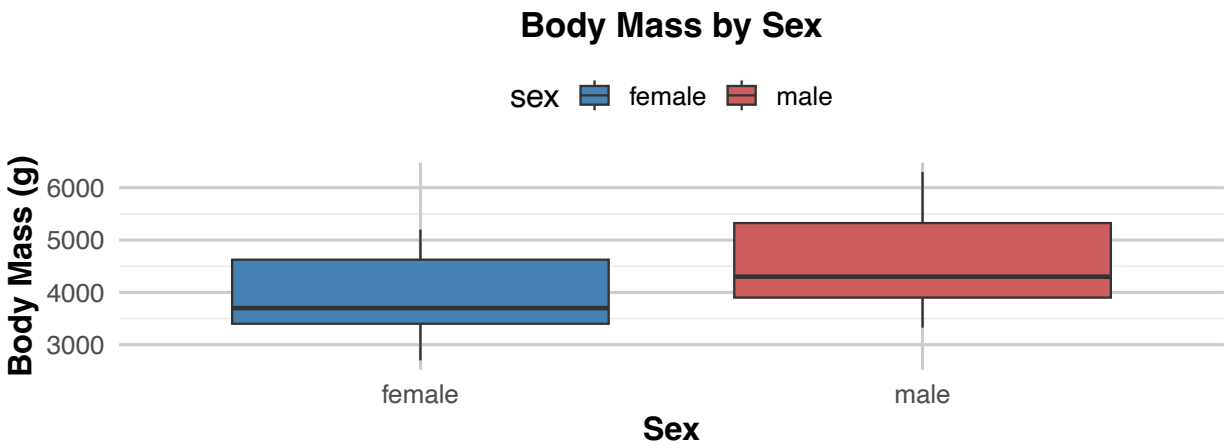
sex	Mean_Body_Mass	Mean_Flipper_Length	Mean_Bill_Length	Mean_Bill_Depth
female	3928.960	198.1485	42.68812	16.33267
male	4534.343	204.8081	45.76263	17.87879

The summary shows that male penguins have a higher average body mass (4534.34 g) and flipper length (204.38 mm) than females, who average 3928.96 g and 198.14 mm, respectively. Males also have slightly

longer bill lengths (45.76 mm) and deeper bills (17.38 mm) compared to females, indicating noticeable differences in physical traits by sex.

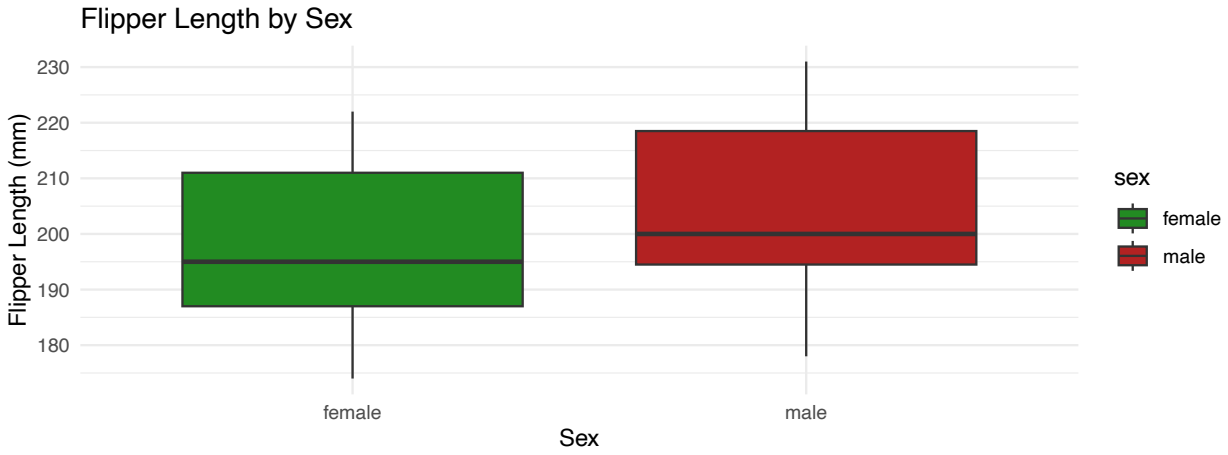
Graphical Analysis

Box Plot of Body Mass by Sex



The box plot shows that male penguins typically have higher body mass than females, with minimal overlap. This difference in body mass between sexes suggests it may be a reliable indicator for sex estimation.

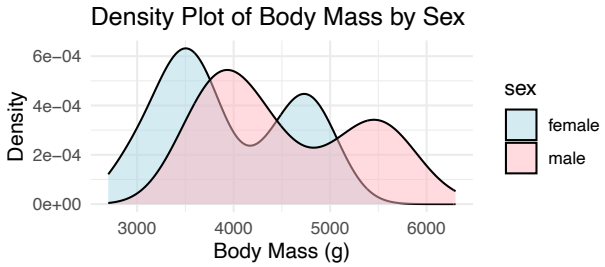
Box Plot of Flipper Length by Sex

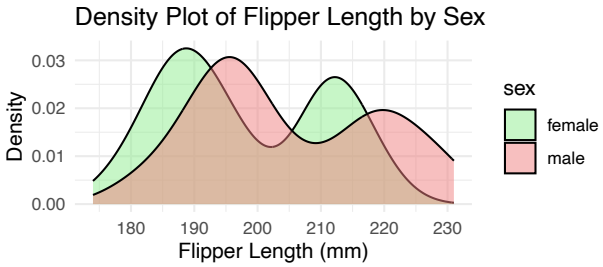


The box plot shows that male penguins generally have longer flippers than females. Males have a median flipper length around 205 mm, with values typically ranging from 190 mm to 230 mm. Female flipper lengths are shorter on average, with a median around 200 mm and a range mostly between 185 mm and 215 mm. This difference suggests that flipper length may be useful in distinguishing between male and female penguins.

Reliability Assessment for Sex Prediction

To assess the reliability of using body mass and flipper length as indicators for sex, we can calculate the separation between male and female measurements using the overlap of distributions and correlation analysis.





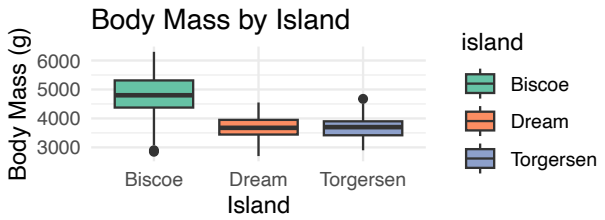
The plots indicate that male penguins typically have higher body mass and longer flipper lengths than females, suggesting both traits are useful for sex prediction, especially body mass.

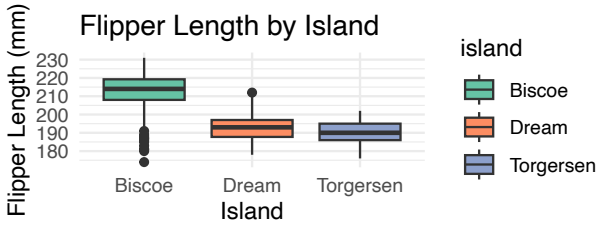
Conclusion

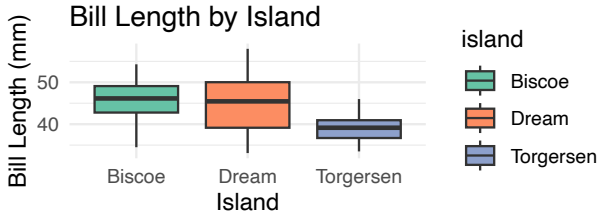
Body mass and flipper length effectively distinguish male from female penguins, with males generally having higher body mass and longer flippers. Body mass shows stronger separation, making it a reliable, non-invasive indicator for sex estimation.

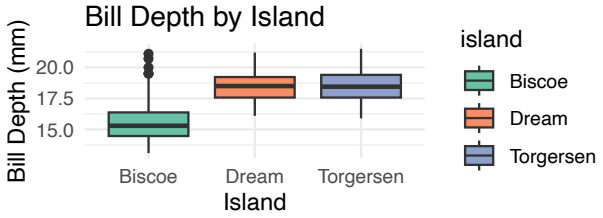
TASK-4: Impact of Island on Penguin Physical Characteristics

The below field plots show the distribution of frame mass, flipper length, invoice length, and invoice intensity with the help of using island to visually examine characteristics.









The boxplots shows that Biscoe penguins have the highest body mass (4,700g), flipper length (210 mm), and bill length (45 mm). Torgersen penguins are lighter (3,700g) with shorter flippers (190 mm) but have the greatest bill depth (18 mm), showing clear island-based differences.

Conclusion

The analysis indicates that island origin significantly affects certain physical characteristics of penguins. ANOVA results show significant differences in body mass ($p < 0.05$) and flipper length ($p < 0.05$) across islands, with median body mass ranging from 3700 g to 5000 g and flipper length from 185 mm to 210 mm depending on the island. These findings suggest that island-specific environmental factors likely influence penguin morphology, potentially due to differences in resources and habitat conditions.