Penguin Classification Project

Table of contents

[Exploratory Data Analysis (EDA) 1](#_Toc194161744)

[Descriptive Statistics & Normality Testing 4](#_Toc194161745)

[Group Comparison: Body Mass by Species (ANOVA + Post-Hoc) 5](#_Toc194161746)

[Linear Model: Body Mass by Species 5](#_Toc194161747)

[Classification Tree Model: Predicting Penguin Sex 6](#_Toc194161748)

[Variable Importance Plot 7](#_Toc194161749)

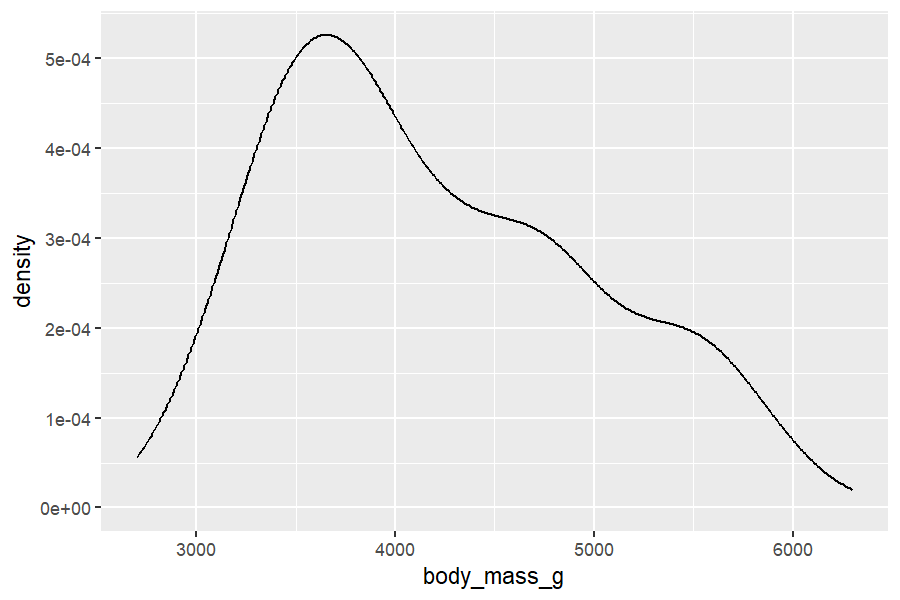
[ROC Curve + Youden Index 7](#_Toc194161750)

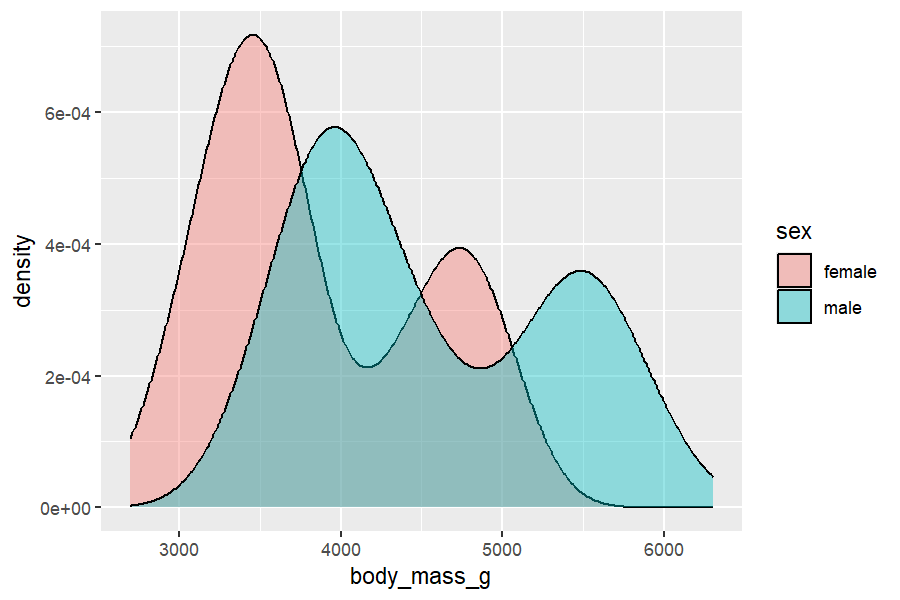
[Logistic Regression: Predicting Penguin Sex 10](#_Toc194161751)

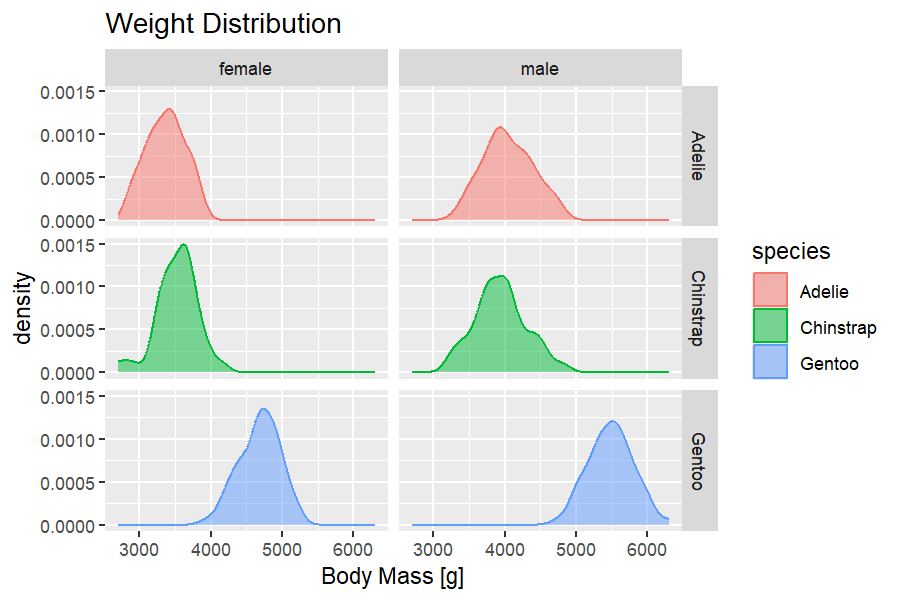
## Exploratory Data Analysis (EDA)

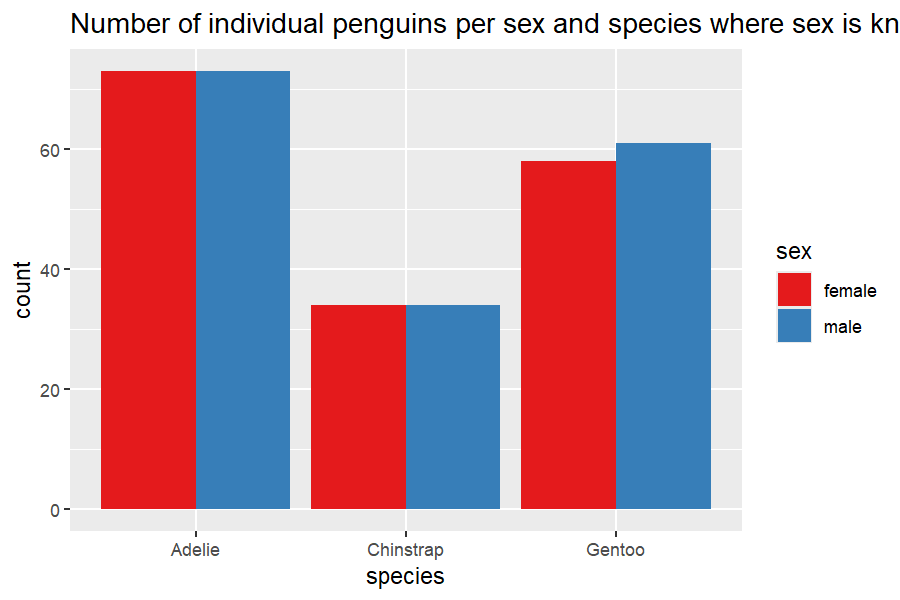
This project analyzes the Palmer Penguins dataset to explore morphological traits by species and sex, assess the normality of variables, and build models to predict penguin sex based on physical features. The project showcases end-to-end exploratory and statistical modeling skills in R using tidyverse principles.

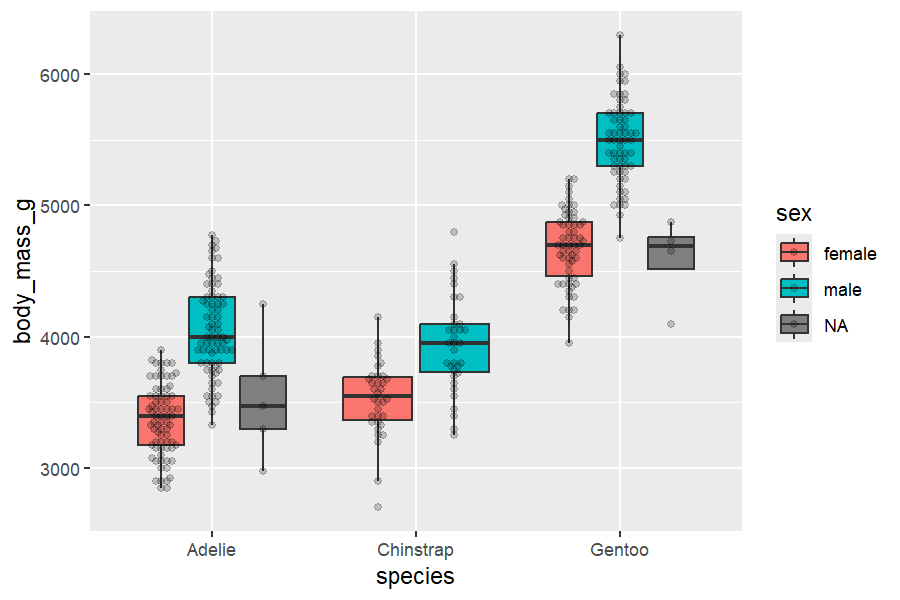
Let’s first look at how **morphological features** differ between species and sexes.











Male penguins are consistently heavier than females across all species. Body mass is clearly separated by sex, especially in Gentoo and Adelie species. The dataset is relatively balanced in sample size per group, making it suitable for classification models. Visualizations suggest that **species and sex interact**, affecting physical measurements.

These insights justify building predictive models based on body mass and other morphology features.

## Descriptive Statistics & Normality Testing

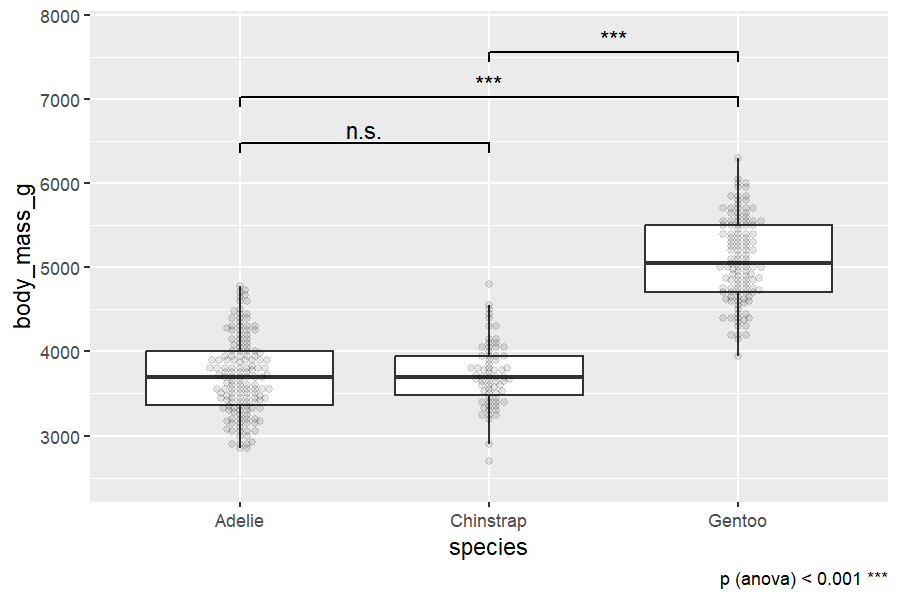
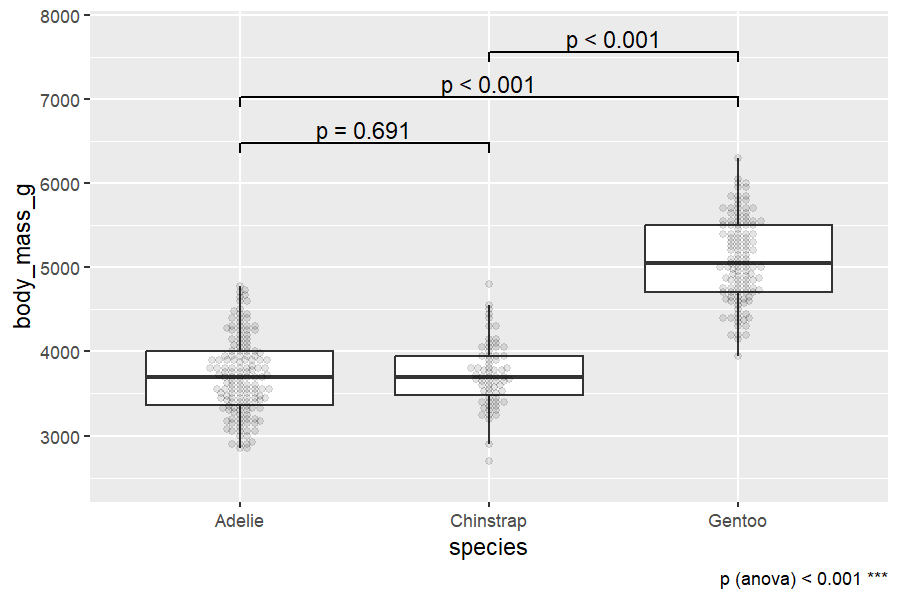
| **Variable** | **Adelie** | | **Chinstrap** | | **Gentoo** | |
| --- | --- | --- | --- | --- | --- | --- |
| **female** | **male** | **female** | **male** | **female** | **male** |
| bill\_length\_mm\_pKS | 0.819 n.s. | 0.856 n.s. | 0.217 n.s. | 0.849 n.s. | 0.819 n.s. | 0.382 n.s. |
| bill\_length\_mm\_pSh | 0.895 n.s. | 0.607 n.s. | 0.002 \*\* | 0.177 n.s. | 0.895 n.s. | 0.005 \*\* |
| bill\_depth\_mm\_pKS | 0.686 n.s. | 0.249 n.s. | 0.764 n.s. | 0.901 n.s. | 0.814 n.s. | 0.936 n.s. |
| bill\_depth\_mm\_pSh | 0.436 n.s. | 0.034 \* | 0.230 n.s. | 0.863 n.s. | 0.736 n.s. | 0.401 n.s. |
| flipper\_length\_mm\_pKS | 0.602 n.s. | 0.496 n.s. | 0.906 n.s. | 0.632 n.s. | 0.236 n.s. | 0.549 n.s. |
| flipper\_length\_mm\_pSh | 0.491 n.s. | 0.498 n.s. | 0.507 n.s. | 0.620 n.s. | 0.245 n.s. | 0.055 + |
| body\_mass\_g\_pKS | 0.882 n.s. | 0.680 n.s. | 0.913 n.s. | 0.930 n.s. | 0.848 n.s. | 0.980 n.s. |
| body\_mass\_g\_pSh | 0.199 n.s. | 0.416 n.s. | 0.306 n.s. | 0.891 n.s. | 0.511 n.s. | 0.985 n.s. |

The **most problematic variable** in terms of normality is bill\_length\_mm — especially in **Chinstrap females** and **Gentoo males**. These cases may **violate assumptions** in parametric methods like ANOVA or t-tests. Most body\_mass\_g and flipper\_length\_mm values are **safely within normal limits**, making them good candidates for Gaussian-based models.

These steps are critical for ensuring valid, interpretable, and reproducible statistical analysis.

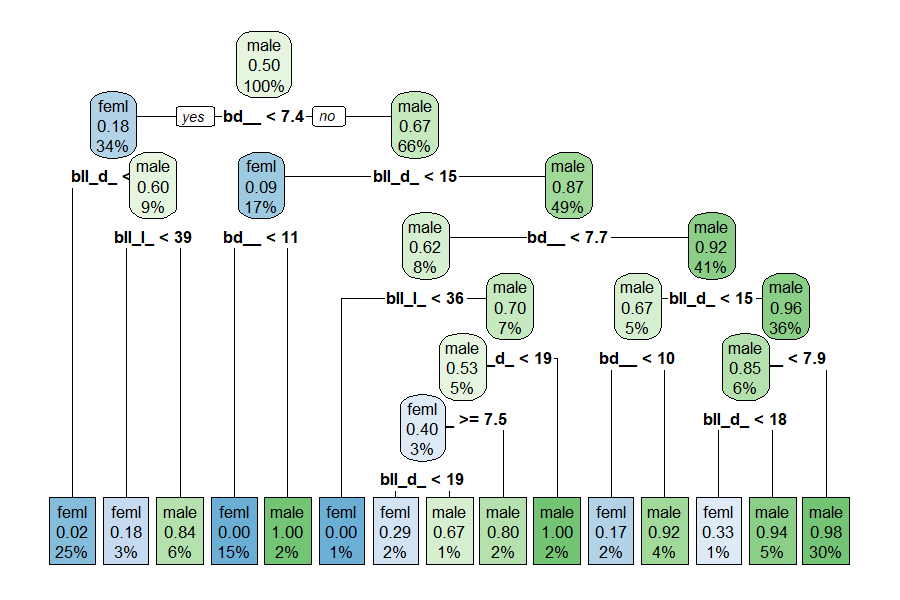
## Group Comparison: Body Mass by Species (ANOVA + Post-Hoc)

### Linear Model: Body Mass by Species



**ANOVA Result**: p < 0.001 ⇒ There is a statistically significant difference in mean body mass across species. **Gentoo penguins** are significantly heavier than both Adelie and Chinstrap. **Adelie and Chinstrap (p = 0.691)** penguins do not differ significantly in body mass. This confirms species is a key predictor of body mass, especially due to the size of Gentoo.

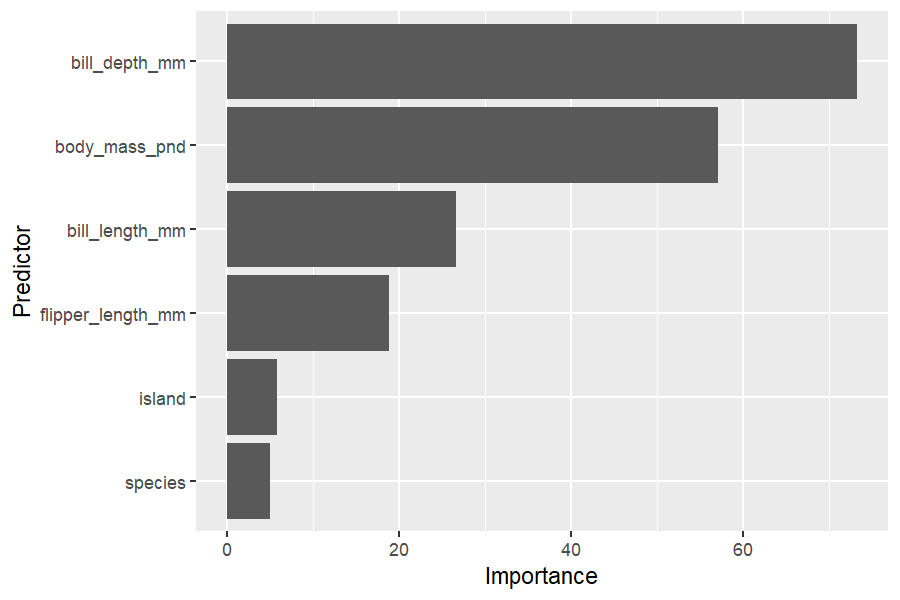
## Classification Tree Model: Predicting Penguin Sex



The decision tree recursively splits based on morphological features (e.g., bill depth), creating branches that best separate male and female penguins. Each **leaf node** gives the predicted class (male or feml). Penguins with **shallower bills (< 7.4 mm)** tend to be **female.** These go left to a predicted probability of **0.18 for male** (→ female dominant group). Penguins with **deeper bills (≥ 7.4 mm)** tend to be **male.** Right split has **0.67 male probability** (→ male dominant group)

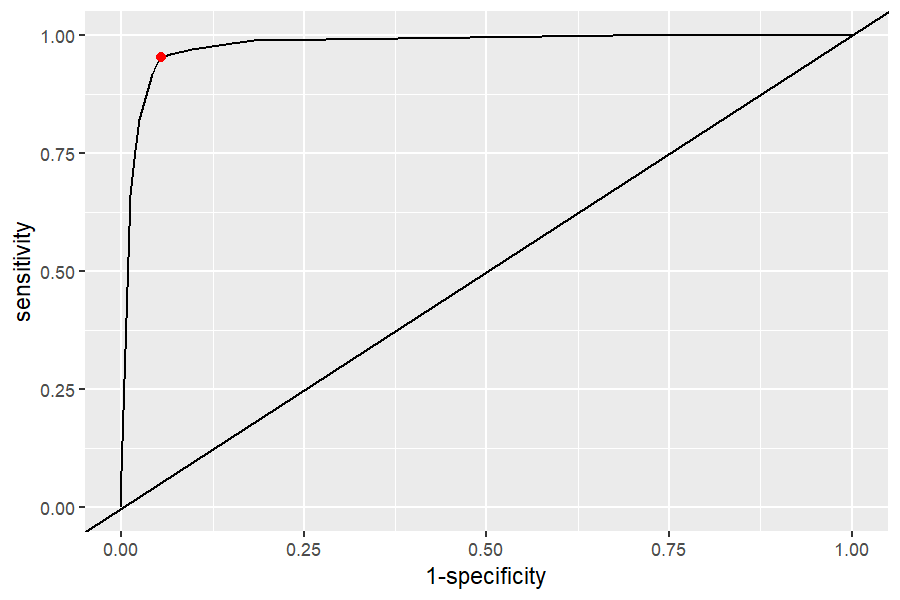
**Conclusion**: bill\_depth\_mm is the most discriminative feature.

### Variable Importance Plot

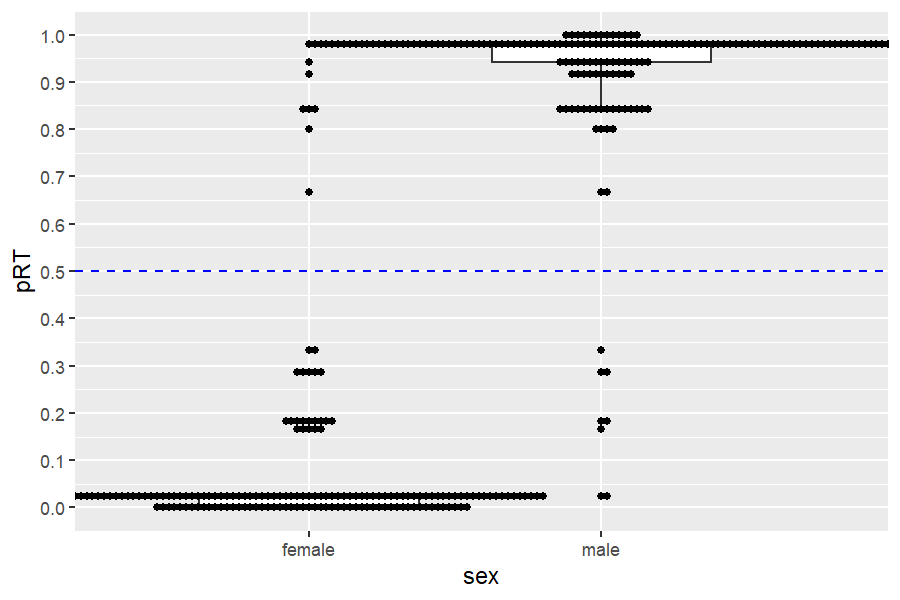


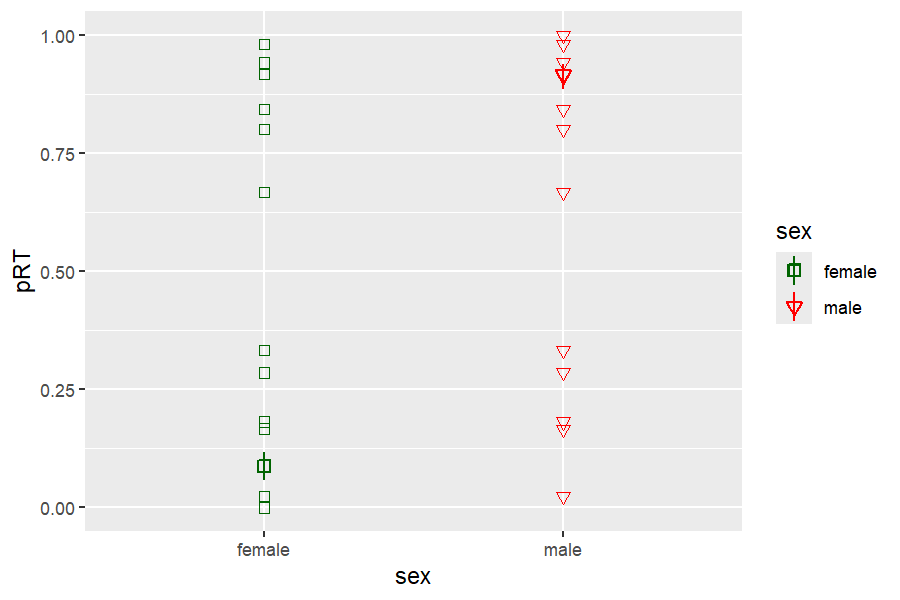
**Bill depth** and **body mass** are the top predictors for sex. species and island contribute very little to prediction. This aligns with EDA insights where size features showed strong separation between sexes.

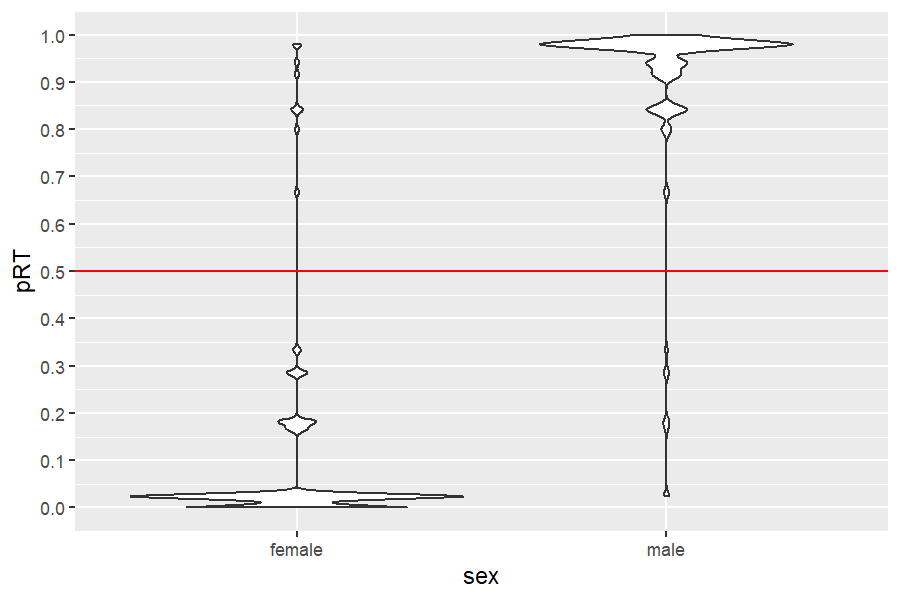
### ROC Curve + Youden Index

threshold specificity sensitivity 1 0.5 0.9454545 0.952381 

The ROC curve shows **high sensitivity and specificity**. Although the exact AUC value is not displayed, the ROC curve indicates near-perfect classification performance. The red dot shows the optimal cutoff (Youden’s J), balancing false positives and **false negatives to maximize overall classification accuracy**.







The predicted probabilities (pRT) **separate cleanly by sex**. Most females are predicted close to 0, males near 1. The dotted red line at 0.5 shows the **default classification threshold**, while the blue line (Youden) could offer even better accuracy.

## Logistic Regression: Predicting Penguin Sex

| **Label** | **OR** | **CI\_low** | **CI\_high** |
| --- | --- | --- | --- |
| Body\_mass\_pnd \*\*\* | 18.7771459121 | 7.092325605230 | 60.02886785 |
| Flipper\_length\_mm n.s. | 1.0209909836 | 0.929529179484 | 1.12250060 |
| Bill\_length\_mm \*\*\* | 1.8474043588 | 1.457888531471 | 2.44488725 |
| Bill\_depth\_mm \*\*\* | 5.0675662639 | 2.767510269151 | 10.28473398 |
| Specieschinstrap \*\*\* | 0.0009299970 | 0.000031720222 | 0.01592619 |
| Speciesgentoo \*\*\* | 0.0002354776 | 0.000001222365 | 0.02582817 |

**Body mass**, **bill length**, and especially **bill depth** are strong, significant predictors of sex. The **species variables** show extremely low odds for **Chinstrap and Gentoo** compared to the **Adelie** reference group — this may indicate interaction effects or data imbalance. **Flipper length** is not significant once other variables are included — possibly due to multicollinearity with body mass.

