Supplemental File

Ryan Grunert

11/26/2021

#Loading required packages  
library(here)

## here() starts at /Users/ryangrunert/Desktop/Fall2021\_Classes/MADA/RyanGrunert-MADA-project

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✓ ggplot2 3.3.5 ✓ purrr 0.3.4  
## ✓ tibble 3.1.6 ✓ dplyr 1.0.7  
## ✓ tidyr 1.1.4 ✓ stringr 1.4.0  
## ✓ readr 2.1.0 ✓ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

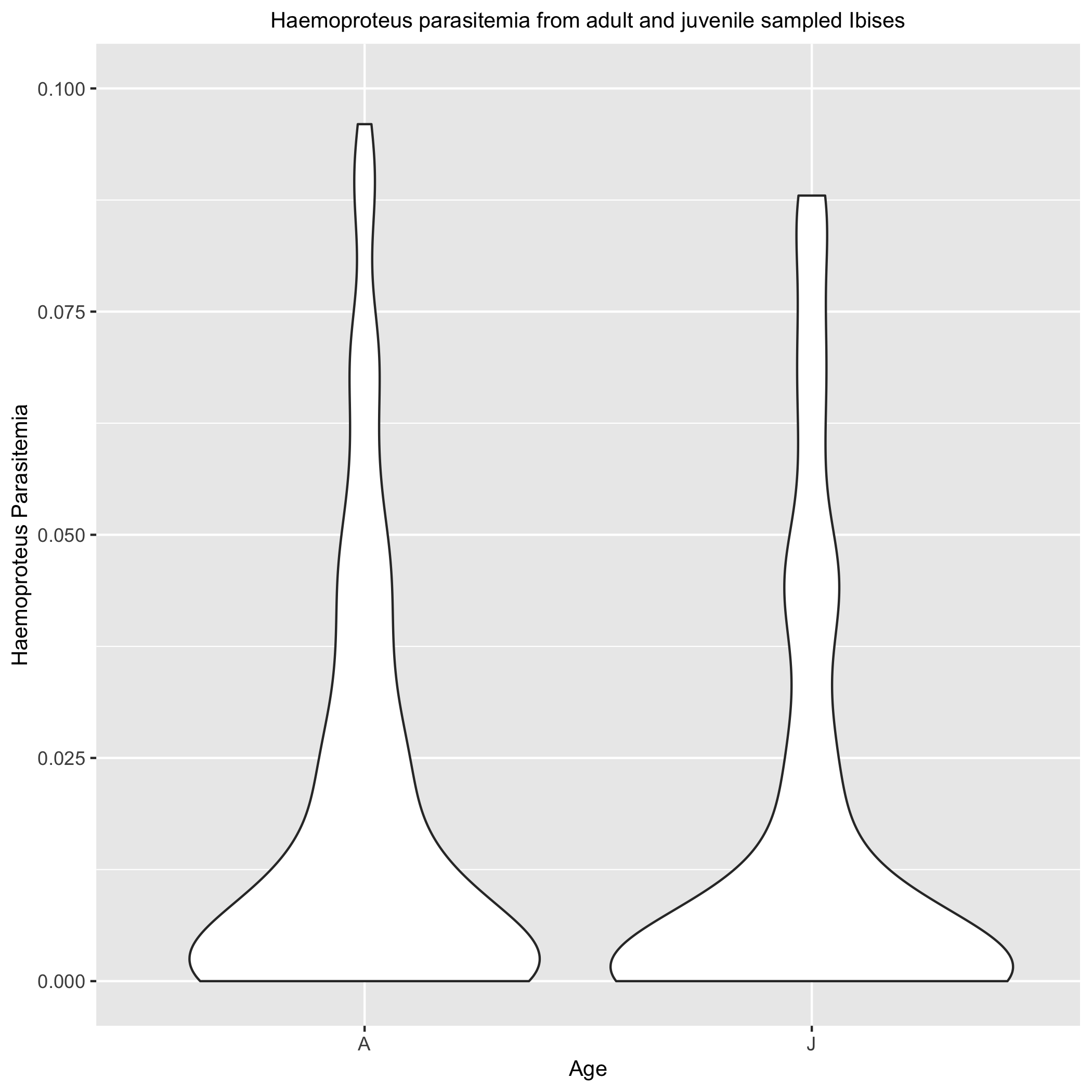
library(knitr)  
#Exploratory Figures  
figure\_Exp\_2 <- here::here("supplemental\_materials", "Figures", "ExpFigure2.png")  
figure\_Exp\_3 <- here::here("supplemental\_materials", "Figures", "ExpFigure3.png")  
figure\_Exp\_5 <- here::here("supplemental\_materials", "Figures", "ExpFigure5.png")  
#Machine Learning Model Diagnostic Figures  
figure\_ML\_1 <- here::here("supplemental\_materials", "Figures", "ImpPlotDT.jpeg")  
figure\_ML\_2 <- here::here("supplemental\_materials", "Figures", "DTTree.jpeg")  
figure\_ML\_3 <- here::here("supplemental\_materials", "Figures", "LassoFig.jpeg")  
#Loading Model Results  
data\_locationPA1 <- here::here("supplemental\_materials", "Tables", "Model1MetricsPA.Rda")  
model1 <- readRDS(data\_locationPA1)  
data\_locationPA2 <- here::here("supplemental\_materials", "Tables", "Model2MetricsPA.Rda")  
model2 <- readRDS(data\_locationPA2)  
data\_locationPA3 <- here::here("supplemental\_materials", "Tables", "Model3MetricsPA.Rda")  
model3 <- readRDS(data\_locationPA3)  
data\_locationPA6 <- here::here("supplemental\_materials", "Tables", "Model6MetricsPA.Rda")  
model6 <- readRDS(data\_locationPA6)  
data\_locationPA7 <- here::here("supplemental\_materials", "Tables", "Model7MetricsPA.Rda")  
model7 <- readRDS(data\_locationPA7)  
data\_locationPA9 <- here::here("supplemental\_materials", "Tables", "Model9MetricsPA.Rda")  
model9 <- readRDS(data\_locationPA9)  
#Loading required materials - Tables  
mod\_loc\_1 <- here::here("results", "Models", "Model1PA.Rda")  
mod\_loc\_2 <- here::here("results", "Models", "Model2PA.Rda")  
mod\_loc\_3 <- here::here("results", "Models", "Model3PA.Rda")  
mod\_loc\_4 <- here::here("results", "Models", "Model4PA.Rda")  
mod\_loc\_5 <- here::here("results", "Models", "Model5PA.Rda")  
mod\_loc\_6 <- here::here("results", "Models", "Model6PA.Rda")  
mod\_loc\_7 <- here::here("results", "Models", "Model7PA.Rda")  
mod\_loc\_8 <- here::here("results", "Models", "Model8PA.Rda")  
mod\_loc\_9 <- here::here("results", "Models", "Model9PA.Rda")

This document details all of the supplemental information and models generated by the project.

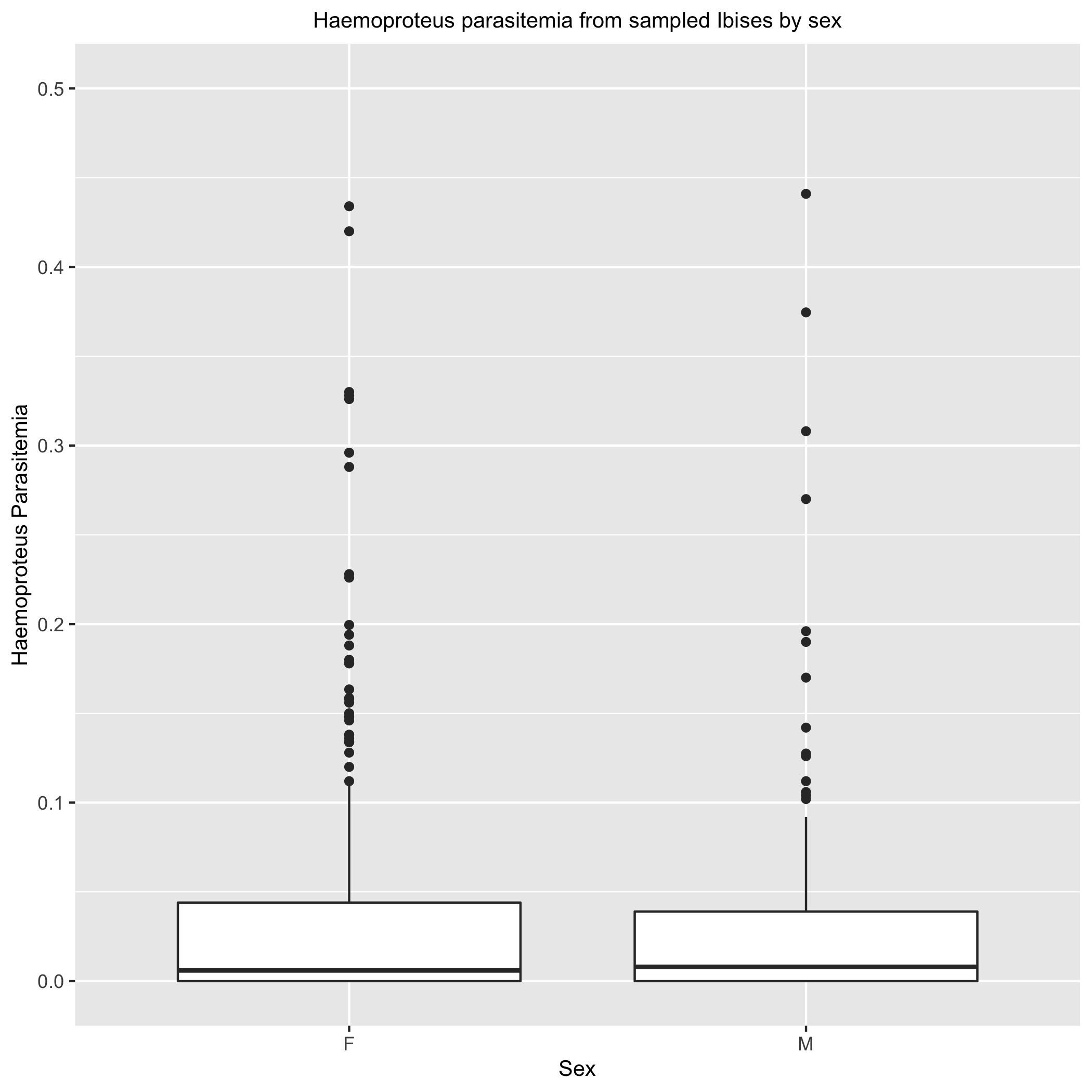
### Exploratory Data Analysis

The following are exploratory data analysis graphs that were considered vital enough to include in the main manuscript.

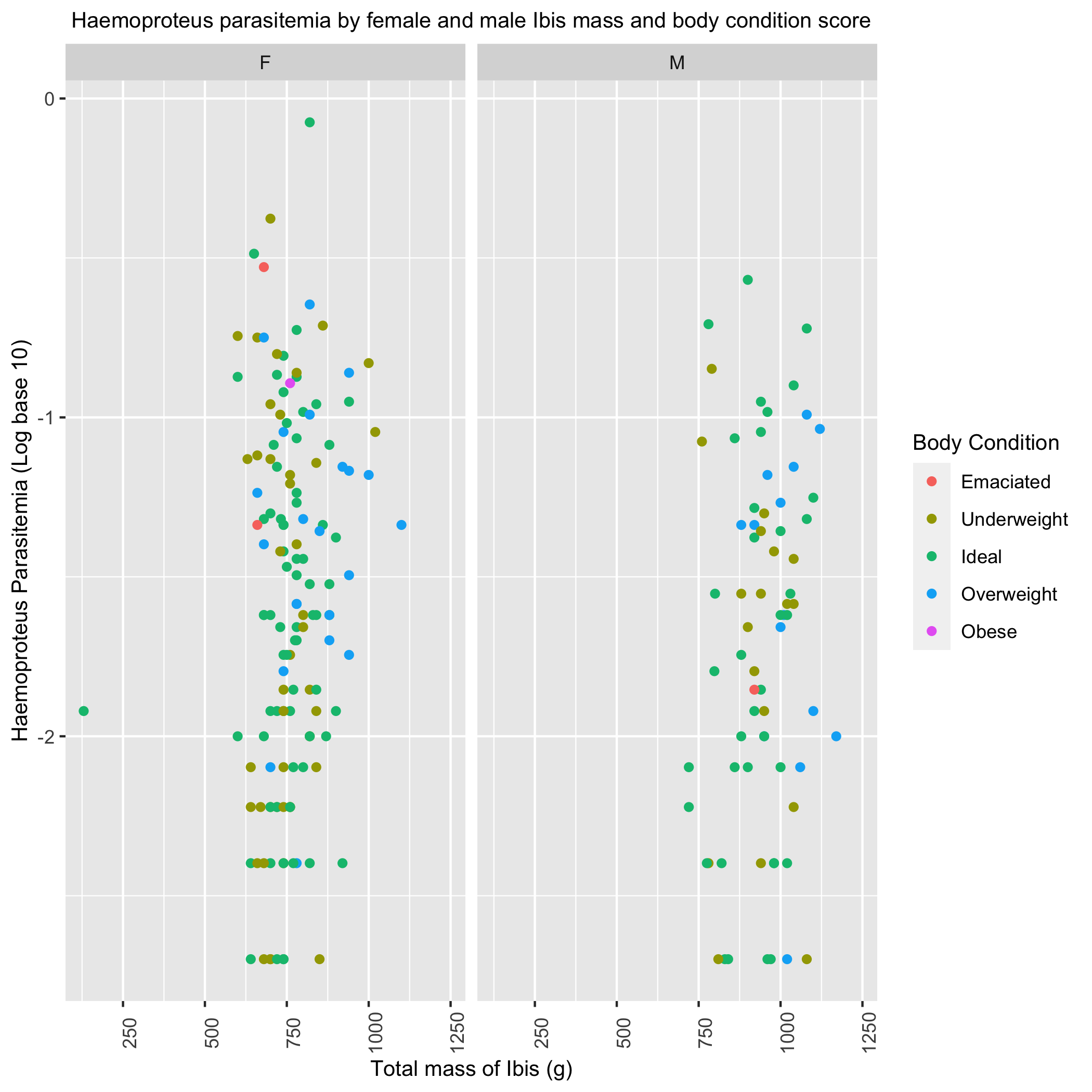
knitr::include\_graphics(figure\_Exp\_2, error = FALSE)

 This is a violin plot of the *Haemoproteus* parasitemia values among infected adult and juvenile white Ibis. As we can see, the majority of the birds have lower parasitemias below 0.025, and a couple outliers that have higher values (most likely due to recent infection, not chronic). The adult birds have a slightly larger spread than the juveniles.

knitr::include\_graphics(figure\_Exp\_3, error = FALSE)

 This figure is a boxplot of the *Haemoproteus* parasitemia distributions of female and male infected Ibis. The females have a slightly larger distribution, but no other clear pattern exists between the sexes.

knitr::include\_graphics(figure\_Exp\_5, error = FALSE)

 This figure is a scatter plot of *Haemoproteus* parasitemia values and individual Ibis body condition stratified by sex. While the male birds do have larger masses, the overall distribution of body conditions is similar between males and females regardless of parasitemia value. This supports the theory that chronic infections with the parasite don’t have a measurable detrimental health effect on the Ibis themselves.

### Univariate Logistic Models

This section details the results of a preliminary statistical analysis on the Haemoproteus parasitemia data just looking at the presence and absence data, as opposed to the amount of parasitemia per sampled bird.

The following models were setup and tested through a 10-fold cross-validation. The training/testing data proportions were 80/20. 10 folds were taken on 80% of the observations, and the results are depicted here.

model1

## # A tibble: 2 × 6  
## .metric .estimator mean n std\_err .config   
## <chr> <chr> <dbl> <int> <dbl> <chr>   
## 1 accuracy binary 0.569 50 0.00985 Preprocessor1\_Model1  
## 2 roc\_auc binary 0.445 50 0.00734 Preprocessor1\_Model1

This univariate classification GLM looked at the effect of age on the presence of Haemoproteus in infected birds. The results were unsatisfactory with an accuracy of 0.569 and roc\_auc of 0.444.

model2

## # A tibble: 2 × 6  
## .metric .estimator mean n std\_err .config   
## <chr> <chr> <dbl> <int> <dbl> <chr>   
## 1 accuracy binary 0.606 50 0.0112 Preprocessor1\_Model1  
## 2 roc\_auc binary 0.534 50 0.0123 Preprocessor1\_Model1

This univariate classification GLM looked at the effect of sex of the Ibis on the presence of Haemoproteus in infected birds. The results were unsatisfactory with an accuracy of 0.606 and roc\_auc of 0.533, the metrics were higher than age.

model3

## # A tibble: 2 × 6  
## .metric .estimator mean n std\_err .config   
## <chr> <chr> <dbl> <int> <dbl> <chr>   
## 1 accuracy binary 0.584 40 0.0102 Preprocessor1\_Model1  
## 2 roc\_auc binary 0.594 40 0.0122 Preprocessor1\_Model1

This univariate classification GLM looked at the effect of sampling site of the Ibis on the presence of Haemoproteus in infected birds. The results were unsatisfactory with an accuracy of 0.583 and roc\_auc of 0.593.

model6

## # A tibble: 2 × 6  
## .metric .estimator mean n std\_err .config   
## <chr> <chr> <dbl> <int> <dbl> <chr>   
## 1 accuracy binary 0.535 50 0.00880 Preprocessor1\_Model1  
## 2 roc\_auc binary 0.476 50 0.00754 Preprocessor1\_Model1

This univariate classification GLM looked at the effect of habitat type on the presence of Haemoproteus in infected birds. The results were unsatisfactory, however, these results were better than the other univariate models so far.

model7

## # A tibble: 2 × 6  
## .metric .estimator mean n std\_err .config   
## <chr> <chr> <dbl> <int> <dbl> <chr>   
## 1 accuracy binary 0.627 50 0.0129 Preprocessor1\_Model1  
## 2 roc\_auc binary 0.511 50 0.0178 Preprocessor1\_Model1

This univariate classification GLM looked at the effect of total bird mass in grams on the presence of Haemoproteus in infected birds. The results were unsatisfactory with an accuracy of 0.626 and roc\_auc of 0.511.

### Multivariate Models

model9

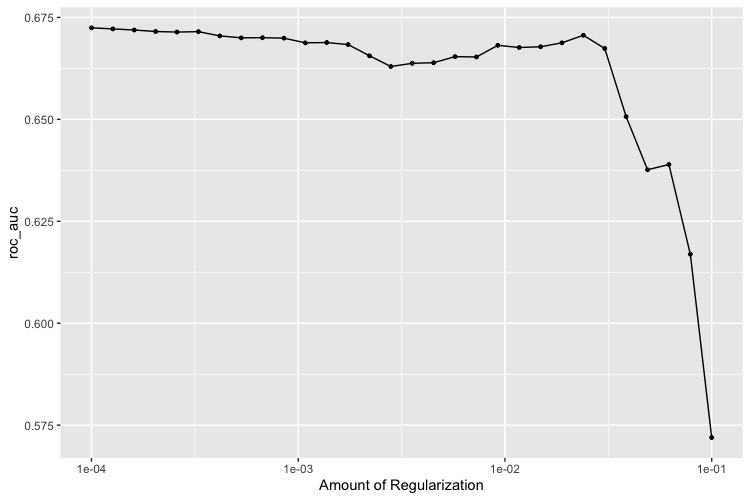
## # A tibble: 2 × 6  
## .metric .estimator mean n std\_err .config   
## <chr> <chr> <dbl> <int> <dbl> <chr>   
## 1 accuracy binary 0.590 40 0.0113 Preprocessor1\_Model1  
## 2 roc\_auc binary 0.608 40 0.0120 Preprocessor1\_Model1

This multivariate classification logistic model looked at the effect of sampling date and site on the presence of *Haemoproteus* in infected birds. The results were unsatisfactory, just a little lower than the previous multivariate classification model. Sampling date and site are important variables to determine the presence of *Haemoproteus*, although season has an effect as well.

### LASSO Model

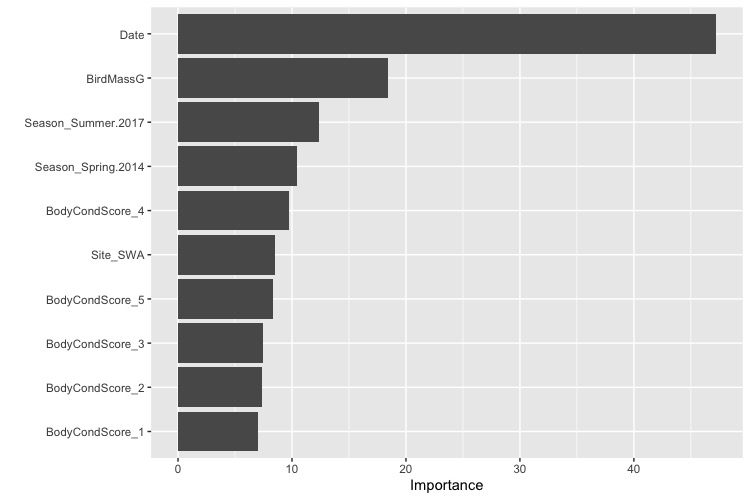
I conducted a Least Absolute Shrinkage and Selection Operator (LASSO) Model with the presence or absence of *Haemoproteus* parasitemia as the outcome variable and using all the predictor variables. The model had an roc\_auc of 0.674. Below is the graph of the regularization.

knitr::include\_graphics(figure\_ML\_3, error = FALSE)



#### Decision Tree Model

The decision tree model was run on all the predictors against the *Haemoproteus* parasitemia presence and absence data. The model was run on all 10 folds of the training data. The roc\_auc of the best tree was 0.681 and the accuracy of the model was 0.631. The two predictors that had the most significant effect on the accuracy of the model were the sample date and mass of the sampled birds in grams. The mos tree with the best parameters had a maximum depth of 8 nodes.

 Looking at the variable importance plot, we can see that sampling date and the mass of the birds in grams are the most significant predictors for the presence of *Haemoproteus* in infected birds. Sampling season is also important, as well as the SWA site. SWA had the highest number of sampled birds.

This concludes the supplemental information for the project.