MADA Project - Ibis Blood Parasites from Florida

Ryan Grunert

2021-10-29

#Loading required packages #for data processing  
library(here) #to set paths

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

library(tidyverse) #all required data manipulation packages

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

## ✓ ggplot2 3.3.5 ✓ purrr 0.3.4  
## ✓ tibble 3.1.5 ✓ dplyr 1.0.7  
## ✓ tidyr 1.1.4 ✓ stringr 1.4.0  
## ✓ readr 2.0.2 ✓ forcats 0.5.1

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

library(xlsx) #to read in excel spreadsheets  
library(janitor) #To assist in cleaning up dates and tidy data

##   
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':  
##   
## chisq.test, fisher.test

library(lubridate) #To extract times from raw data

##   
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':  
##   
## date, intersect, setdiff, union

# 1 Introduction

## 1.1 General Background Information

The data I will be using for this project is American White Ibis (*Eudocimus albus*) blood parasite data that was collected from 2010-2017. The parasite the data focuses on is the protozoan *Haemoproteus*, which causes a chronic blood infection in White Ibises and many other birds. *Haemoproteus* is a vector-borne disease, it is transmitted by *Culicoides* biting midges when they feed on the birds and causes long-lasting parasitemia. Adult healthy birds tolerate haemosporidian infections well, but young birds and birds outside their normal range that are naive to the pathogen are more likely to develop clinical disease. Haemosporidian infections usually cause long-lasting chronic infections in natural hosts, and can cause reduced reproductive success, host fitness, and increased stress and disease susceptibility.

Here is a link to a previously published study concerning this data -> <https://www.sciencedirect.com/science/article/pii/S221322441730072X>

## 1.2 Description of data and data source

The data was collected in a collaboration between Dr. Michael Yabsley’s lab (of which I am a part) and Dr. Sonia Hernandez’s lab here at UGA. I received the raw data from Dr. Yabsley in an agreement to analyze the data over the course of the semester in exchange for using the data for the MADA project.

The data is organized in excel spreadsheets. The two main spreadsheets (Ibis\_Blood\_ParasiteData\_2017 and IbisFieldDataOct19\_2017) hold the majority of the data, and two supplemental spreadsheets from other collaborators that were or currently working on the project contain additional variables that will be combined with the main dataset in R for additional analyses. Each observation in the data is an individual White Ibis that was captured and sampled, and each variable is a different descriptor of either the Ibis itself, or the site/environment it was captured in.

Two datasheets have overlapping information (Ibis\_Blood\_ParasiteData\_2017 and IbisFieldDataOct19\_2017), but both will be combined to have a complete dataset for analysis. The first contains parasitemia data, while the second contains field data. The 100IbisID\_UrbanGradient datasheet will be used to assist with the analysis concerning *Haemoproteus* presence and habitat type. The SalmonellaFieldSerotypes datasheet may be used to compare the presence of both *Haemoproteus* and *Salmonella* in the Ibis, although that may not be needed considering the difference in biology and epidemiology of the two pathogens.

#Setting the path to the Ibis blood parasite spreadsheet  
data\_locationIbisB <- here::here("raw\_data","Ibis\_Blood\_ParasiteData\_2017.xlsx")  
  
#Loads the raw Ibis blood data excel sheet from 2015-2017  
rdIbisBlood15\_17 <- read.xlsx(data\_locationIbisB, sheetName = "2015-2017")  
  
#Loads the raw Ibis blood data excel sheet from 2010-2014  
rdIbisBlood10\_14 <- read.xlsx(data\_locationIbisB, sheetName = "2010-2014")  
  
str(rdIbisBlood15\_17)

## 'data.frame': 406 obs. of 42 variables:  
## $ X. : chr "NSFJB01" "NSFJB02" "NSFJB03" "NSFJB04" ...  
## $ Anje.needs.ASAP : num NA NA NA NA NA NA NA NA NA NA ...  
## $ NA. : num NA NA NA NA NA NA NA NA 1 NA ...  
## $ sex.needed.for.microbiome : chr NA NA NA NA ...  
## $ Haem.PCR.result : chr NA "0" "0" "0" ...  
## $ Haem.seq : chr NA NA NA NA ...  
## $ PCR.sex : chr "not bled" "F" "F" "M" ...  
## $ Slide.read........reader.init : chr NA NA NA NA ...  
## $ F1 : num NA NA NA NA NA 657 NA NA NA NA ...  
## $ F2 : num NA NA NA NA NA 549 NA NA NA NA ...  
## $ F3 : num NA NA NA NA NA 611 NA NA NA NA ...  
## $ Avg.RBC.FOV : num 0 0 0 0 0 ...  
## $ X.FOV.to.read : chr "#DIV/0!" "#DIV/0!" "#DIV/0!" "#DIV/0!" ...  
## $ Total.RBC : chr "#DIV/0!" "50000" "50000" "50000" ...  
## $ X..Haemoproteus : num NA 2 19 8 0 15 3 1 0 4 ...  
## $ Haemoproteus.parasitemia. : chr "#DIV/0!" "0.004" "0.038" "0.016" ...  
## $ Comments..e.g...young.forms.present. : chr NA NA NA NA ...  
## $ Blood.Smears..2. : chr "N" "Y (after hep)" "Y" "Y" ...  
## $ Date : chr "42289" "42289" "42289" "42289" ...  
## $ Season : chr "Fall 2015" "Fall 2015" "Fall 2015" "Fall 2015" ...  
## $ NA..1 : chr NA NA NA NA ...  
## $ Site.Name : chr "Juno Beach" "Juno Beach" "Juno Beach" "Juno Beach" ...  
## $ Habitat.type : chr "Urban" "Urban" "Urban" "Urban" ...  
## $ Ibis.. : num 1 2 3 4 5 5 6 7 8 9 ...  
## $ Age : chr "Juvenile" "Adult" "Adult" "Adult" ...  
## $ Total.mass..g. : chr "900" "820" "840" "910" ...  
## $ Mass.bag..g. : chr "100" "50" "100" "112" ...  
## $ Mass.bird..g. : chr "800" "770" "740" "798" ...  
## $ Body.condition.score..1.5. : chr "2" "3" "3" "3" ...  
## $ Ectoparasite.score..1.5. : chr "1" "2" "2" "1" ...  
## $ Culmen.length..mm. : chr "152.335" "120.59" "123.44" "152" ...  
## $ Wing.chord.length..mm. : chr "291" "278" "261" "297" ...  
## $ Tarsus.length..mm. : chr "102.13" "89.87" "88.36" "97.44" ...  
## $ Tarsus.width..mm. : chr "9.435" "9.52" "9.39" "10.45" ...  
## $ Volume.serology.bleed.B.and.C..ul..after.bka.and.cort.: chr "-999" "300" "420" "40" ...  
## $ NA..2 : logi NA NA NA NA NA NA ...  
## $ NA..3 : logi NA NA NA NA NA NA ...  
## $ SCWDS.box.number : num NA 1 NA 1 1 3 1 NA 1 1 ...  
## $ NA..4 : logi NA NA NA NA NA NA ...  
## $ NA..5 : logi NA NA NA NA NA NA ...  
## $ NA..6 : logi NA NA NA NA NA NA ...  
## $ NA..7 : logi NA NA NA NA NA NA ...

str(rdIbisBlood10\_14)

## 'data.frame': 292 obs. of 36 variables:  
## $ X. : chr "1" "2" "AN330" "AN4" ...  
## $ Site.ID : chr "Royal Palms" "Royal Palms" "Everglades" "Everglades" ...  
## $ Date : num 2010 2010 NA NA NA ...  
## $ PCR.sex : chr NA NA NA NA ...  
## $ County : logi NA NA NA NA NA NA ...  
## $ Time.of.Capture : chr NA NA NA NA ...  
## $ Age..adult.juvenile. : chr NA NA NA NA ...  
## $ Weight.in.pillowcase..g : num NA NA NA NA NA NA NA NA NA NA ...  
## $ Weight.of.pillowcase..g : num NA NA NA NA NA NA NA NA NA NA ...  
## $ Weight.of.bird..g. : num NA NA NA NA NA NA NA NA NA NA ...  
## $ Time.of.blood.collection : chr NA NA NA NA ...  
## $ Body.condition.score..out.of.3 : num NA NA NA NA NA NA NA NA NA NA ...  
## $ Ectoparasite.score..out.of.5 : num NA NA NA NA NA NA NA NA NA NA ...  
## $ Culmen.length..cm. : num NA NA NA NA NA NA NA NA NA NA ...  
## $ Wing.chord..cm. : num NA NA NA NA NA NA NA NA NA NA ...  
## $ Tibia.tarsus.length..cm. : num NA NA NA NA NA NA NA NA NA NA ...  
## $ Tibia.tarsus.width..mm. : num NA NA NA NA NA NA NA NA NA NA ...  
## $ Fecal.sample. : chr NA NA NA NA ...  
## $ Blood.sample. : chr NA NA NA NA ...  
## $ Physiological.Comments : chr NA NA NA NA ...  
## $ Bleeding.notes : chr NA NA NA NA ...  
## $ General.notes : chr NA NA NA NA ...  
## $ Blood.Smear. : chr NA NA NA NA ...  
## $ Blood.for.malaria...SMC.received...in.SCWDS.blue.: num NA NA NA NA NA 1 1 1 1 1 ...  
## $ PCR.malaria..1.0....SMC : num 0 0 0 0 0 1 1 1 1 0 ...  
## $ Sequenced.malaria...SMC : chr "0" "0" "0" "0" ...  
## $ Leuco.SMC : num 0 0 0 0 0 NA NA NA NA NA ...  
## $ Slide.read........reader.init : chr NA NA NA NA ...  
## $ Avg.RBC.FOV : num NA NA NA NA NA ...  
## $ X.FOV : num NA NA NA NA NA 41 27 33 32 82 ...  
## $ Total.RBC : num 5000 5000 NA NA NA ...  
## $ X..Haemoproteus : num 0 0 NA NA NA 102 9 8 40 0 ...  
## $ Scan.HP..y.n. : chr NA NA NA NA ...  
## $ Haemoproteus.parasitemia. : chr "0" "0" NA NA ...  
## $ X..Plasmodium : num NA NA NA NA NA NA NA NA NA NA ...  
## $ Scan.PM..y.n. : chr NA NA NA NA ...

The main variables I’m interested in for these two sheets above are “Haemoproteus.parasitemia.” and the first “X” column is the sample/bird ID. The 2010-14 set is missing more data, but characteristics of the birds from that group may be used.

#Setting the path to the Ibis field data spreadsheet  
data\_locationIbisF <- here::here("raw\_data","IbisFieldDataOct19\_2017.xlsx")  
  
#Loads the raw Ibis field data excel sheet  
rdIbisField <- read.xlsx(data\_locationIbisF, sheetName = "All capture data")  
  
str(rdIbisField)

## 'data.frame': 476 obs. of 94 variables:  
## $ ID..NSFSITE... : chr "-999" "-999" "-999" "-999" ...  
## $ Date : Date, format: "2015-10-12" "2015-10-20" ...  
## $ Season : chr "Fall 2015" "Fall 2015" "Fall 2015" "Fall 2015" ...  
## $ Site.Name : chr "Juno Beach" "J.W. Corbett Wildlife Management Area" "Loxahatchee Wildlife Refuge" "Loxahatchee Wildlife Refuge" ...  
## $ Habitat.type : chr "Urban" "Natural" "Natural" "Natural" ...  
## $ Latitude : num 26.9 26.8 26.5 26.5 26.5 ...  
## $ Longitude : num -80.1 -80.3 -80.2 -80.2 -80.2 ...  
## $ Time.weather.start : chr "0.576388888888889" "0.309722222222222" "0.298611111111111" "0.695833333333333" ...  
## $ Temperature.start : chr "82.6" "74.5" "74.8" "82.9" ...  
## $ Relative.humidity.start : chr "66.5" "72.6" "81.5" "64.7" ...  
## $ Heat.stress.index.start : chr "88.5" "75.8" "76.1" "88.9" ...  
## $ Wind.start : chr "1.2" "0" "0.6" "2.5" ...  
## $ Time.weather.end : chr "NA" "NA" "0.395833333333333" "0.805555555555555" ...  
## $ Temperature.end : chr "NA" "NA" "81" "78.5" ...  
## $ Relative.humidity.end : chr "NA" "NA" "69.5" "77.9" ...  
## $ Heat.stress.index.end : chr "NA" "NA" "85.6" "83.1" ...  
## $ Wind.end : chr "NA" "NA" "1.3" "0.7" ...  
## $ Precipitation : chr "NA" "NA" "N" "N" ...  
## $ Ibis.number : num -999 -999 -999 -999 -999 -999 -999 -999 -999 -999 ...  
## $ Net.opened...Arrived.for.leg.lasso : chr "0.569444444444444" "0.28125" "0.291666666666667" "0.6875" ...  
## $ Net.closed...Stopped.leg.lasso : chr "0.590277777777778" "0.330555555555556" "0.368055555555556" "0.802083333333333" ...  
## $ Capture.effort..minutes. : chr "30.0000000000001" "71" "110" "165" ...  
## $ Time.captured : chr "No Captures" "No Captures" "No Captures" "No Captures" ...  
## $ Recorded.by : chr "-999" "-999" "-999" "-999" ...  
## $ Capture.method : chr "-999" "-999" "-999" "-999" ...  
## $ Age : chr "-999" "-999" "-999" "-999" ...  
## $ Mass.bird..g. : chr "-999" "-999" "-999" "-999" ...  
## $ Body.condition.score..1.5. : chr "-999" "-999" "-999" "-999" ...  
## $ Ectoparasite.score..1.5. : chr "-999" "-999" "-999" "-999" ...  
## $ Culmen.length..mm. : chr "-999" "-999" "-999" "-999" ...  
## $ Wing.chord.length..mm. : chr "-999" "-999" "-999" "-999" ...  
## $ Tarsus.length..mm. : chr "-999" "-999" "-999" "-999" ...  
## $ Tarsus.width..mm. : chr "-999" "-999" "-999" "-999" ...  
## $ PCR.Sex : chr "-999" "-999" "-999" "-999" ...  
## $ Band.Code : chr "-999" "-999" "-999" "-999" ...  
## $ GPS.tagged..over.866g.with.vhf..over.776.6g.966.without.vhf. : chr "-999" "-999" "-999" "-999" ...  
## $ GPS.ID : num -999 -999 -999 -999 -999 -999 -999 -999 -999 -999 ...  
## $ VHF.Function : chr "-999" "-999" "-999" "-999" ...  
## $ VHF.ID : chr "-999" "-999" "-999" "-999" ...  
## $ VHF.Frequency : chr "-999" "-999" "-999" "-999" ...  
## $ Feces.collected.for.general.use..fec..or.stable.isotope.analysis..FS..aim.for.1g. : chr "-999" "-999" "-999" "-999" ...  
## $ Mass.feces.collected..g. : chr "-999" "-999" "-999" "-999" ...  
## $ Feces.collected.for.corticosterone..FC..aim.for.1g. : chr "-999" "-999" "-999" "-999" ...  
## $ Mass.feces.corticosterone..g. : chr "-999" "-999" "-999" "-999" ...  
## $ Time.fecal.sample : chr "-999" "-999" "-999" "-999" ...  
## $ Fecal.samples.for.salmonella..media. : chr "-999" "-999" "-999" "-999" ...  
## $ Salmonella.status.CHROMagar.confirmed : num -999 -999 -999 -999 -999 -999 -999 -999 -999 -999 ...  
## $ Choanal.swab..plastic..purple.bullet. : chr "-999" "-999" "-999" "-999" ...  
## $ Cloacal.swab..wood..green.bullet. : chr "-999" "-999" "-999" "-999" ...  
## $ Blood.Smears..2. : chr "-999" "-999" "-999" "-999" ...  
## $ Smears.by : chr "-999" "-999" "-999" "-999" ...  
## $ Formalin. : chr "-999" "-999" "-999" "-999" ...  
## $ Feather.collected. : chr "-999" "-999" "-999" "-999" ...  
## $ Body.feathers..3. : chr "-999" "-999" "-999" "-999" ...  
## $ Time.bled.A : chr "-999" "-999" "-999" "-999" ...  
## $ Volume.bleed.A..ml. : chr "-999" "-999" "-999" "-999" ...  
## $ Bled.by : chr "-999" "-999" "-999" "-999" ...  
## $ Time.bled.B : chr "-999" "-999" "-999" "-999" ...  
## $ Volume.bleed.B..ml. : chr "-999" "-999" "-999" "-999" ...  
## $ Bled.by.1 : chr "-999" "-999" "-999" "-999" ...  
## $ Time.C : chr "-999" "-999" "-999" "-999" ...  
## $ Volume.bleed.C..ml. : chr "-999" "-999" "-999" "-999" ...  
## $ Bled.by.2 : chr "-999" "-999" "-999" "-999" ...  
## $ Time.spun : chr "-999" "-999" "-999" "-999" ...  
## $ Volume.BKA.bleed.A..ul..min.60. : chr "-999" "-999" "-999" "-999" ...  
## $ Volume.corticosterone.bleed.A..ul..priority...min.60. : chr "-999" "-999" "-999" "-999" ...  
## $ Volume.BKA.bleed.B..ul..min.60. : chr "-999" "-999" "-999" "-999" ...  
## $ Volume.corticosterone.bleed.B..ul..priority...min.60. : chr "-999" "-999" "-999" "-999" ...  
## $ Volume.corticosterone.bleed.C..ul..priority...min.60. : chr "-999" "-999" "-999" "-999" ...  
## $ Volume.serology.bleed.B.and.C..ul..after.bka.and.cort. : chr "-999" "-999" "-999" "-999" ...  
## $ Plasma.biochemistry..ul. : num -999 -999 -999 -999 -999 -999 -999 -999 -999 -999 ...  
## $ d15N : num -999 -999 -999 -999 -999 -999 -999 -999 -999 -999 ...  
## $ d13C : num -999 -999 -999 -999 -999 -999 -999 -999 -999 -999 ...  
## $ Anthropogenic : num -999 -999 -999 -999 -999 -999 -999 -999 -999 -999 ...  
## $ Landfill : num -999 -999 -999 -999 -999 -999 -999 -999 -999 -999 ...  
## $ Fiddler.crabs : num -999 -999 -999 -999 -999 -999 -999 -999 -999 -999 ...  
## $ Fish : num -999 -999 -999 -999 -999 -999 -999 -999 -999 -999 ...  
## $ FW\_Terr\_Inverts : num -999 -999 -999 -999 -999 -999 -999 -999 -999 -999 ...  
## $ Behaviour.of.sampled.ibis..Aggression.during.handling..stressed..in.a.flock.or.alone..etc..: chr "-999" "-999" "-999" "-999" ...  
## $ In.a.flock....2.WHIB.. : chr "-999" "-999" "-999" "-999" ...  
## $ Flock.size.captured.ibis : chr "-999" "-999" "-999" "-999" ...  
## $ Habituation.score.of.flock..1.5. : chr "-999" "-999" "-999" "-999" ...  
## $ Largest.number.ibis.present : chr "-999" "-999" "-999" "-999" ...  
## $ Area.covered.by.flock..m.2. : chr "-999" "-999" "-999" "-999" ...  
## $ Highest.ibis.density : chr "-999" "-999" "-999" "-999" ...  
## $ Number.of.juveniles.in.flock : chr "-999" "-999" "-999" "-999" ...  
## $ Other.species.present.in.flock..names. : chr "-999" "-999" "-999" "-999" ...  
## $ Other.species.present..abundance. : chr "-999" "-999" "-999" "-999" ...  
## $ Number.people.present..largest.group. : chr "-999" "-999" "-999" "-999" ...  
## $ Ibis.observed.eating.human.food. : chr "-999" "-999" "-999" "-999" ...  
## $ People.observed.feeding.ibis. : chr "-999" "-999" "-999" "-999" ...  
## $ Ibis.observed.eating.natural.food : chr "-999" "-999" "-999" "-999" ...  
## $ Obvious.disturbance.at.site. : chr "-999" "-999" "-999" "-999" ...  
## $ Notes : chr NA "3 nets - 1 set up by pin and 2 in v further south on roost site of road. Jeff sent us dropped pin at location o"| \_\_truncated\_\_ "3 nets - one across burm and 2 in v off of diagonal walking trail" NA ...

This dataset is the most organized sheet, but columns of interest will be extracted from this sheet as well as the others to create new clean data frames for analysis.

#Setting the path to the first Ibis supplemental excel sheet (Urban Gradients)  
data\_locationIbisUr <- here::here("raw\_data","100IbisID\_UrbanGradient.xlsx")  
  
#Loads the raw Ibis Urban Gradient supplemental data sheet  
rdUrbanIbis <- read.xlsx(data\_locationIbisUr, sheetName = "Samples by Date")  
  
str(rdUrbanIbis)

## 'data.frame': 100 obs. of 14 variables:  
## $ Name : chr "NSFJB05" "NSFIC02" "NSFIC07" "NSFIC03" ...  
## $ Collection.Date : Date, format: "2015-10-12" "2015-10-13" ...  
## $ Weight : num 0.0137 NA NA NA NA NA NA NA NA NA ...  
## $ mg.kg.Hg..PPM. : num 6.73 NA NA NA NA ...  
## $ Site.Name : chr "Juno Beach" "Indian Creek" "Indian Creek" "Indian Creek" ...  
## $ Site...Urbanized: num 0.527 0.676 0.676 0.676 0.437 ...  
## $ Serotype : chr NA "Rubislaw" "Uganda" NA ...  
## $ NA. : chr NA "KEY" "Subgroup 1: Urbanization (0.00 - 33.8%)| Salmonella + " "Subgroup 2: Urbanization (0.00 - 33.8%)| Salmonella - " ...  
## $ NA..1 : logi NA NA NA NA NA NA ...  
## $ NA..2 : num NA NA NA NA NA NA NA 28 53 19 ...  
## $ NA..3 : logi NA NA NA NA NA NA ...  
## $ NA..4 : logi NA NA NA NA NA NA ...  
## $ NA..5 : chr NA NA "Total # Sites: 6" "Total # Sites: 9" ...  
## $ NA..6 : logi NA NA NA NA NA NA ...

These samples from this datasheet will be used in conjunction with the two main datasheets in order to create an “urbanization gradient” for the samples that are located in either rural or urban areas. The gradient will most likely range from least (“1,” rural) to most (“4,” urban) urban for example, but further work and planning needs to be done first in order to figure out the specifics of the analysis.

#Setting the path to the second Ibis supplemental excel sheet (Salmonella)  
data\_locationIbisSal <- here::here("raw\_data","SalmonellaFieldSerotypesAug27\_2018.xlsx")  
  
#Loads the raw Ibis Urban Gradient supplemental data sheet  
SalIbis <- read.xlsx(data\_locationIbisSal, sheetName = "ALL Positives")  
  
str(SalIbis)

## 'data.frame': 150 obs. of 10 variables:  
## $ Sample.Period : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ Sample.I.D. : chr "NSFIC02FSAL" "NSFIC07FSAL" "NSFDH02FSAL" "NSFDH03FSAL" ...  
## $ Source : chr "fecal" "fecal" "fecal" "fecal" ...  
## $ Season : chr "Fall" "Fall" "Fall" "Fall" ...  
## $ Site.Type : chr "Urban" "Urban" "Urban" "Urban" ...  
## $ Location : chr "Indian Creek" "Indian Creek" "Dreher Park" "Dreher Park" ...  
## $ Days.Before.Processed: num 4 2 5 5 5 5 4 4 4 4 ...  
## $ Stored.on.LB.Stab : num 2 2 2 1 2 2 2 2 2 2 ...  
## $ Positive : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ Serotype : chr "Rubislaw" "Uganda" "Baildon" "Montevideo" ...

This sheet details the Ibises that tested positive for Salmonella, but this data may not be used because of the differences in pathogenesis and epidemiology of the two diseases.

## 1.3 Questions/Hypotheses to be addressed

The questions I plan to address are whether the amounts of *Haemoproteus* parasitemia in the various Ibises are significantly different when focusing on age, sex, season, location, or habitat type (Urban and Rural). Does the amount of parasitemia in the birds decrease, increase or remain the same over time as they get older? Is parasitemia significantly higher in birds sampled in urban or rural environments? These are the main variables I will be looking at for this project, but I may analyze others as the project continues. Not all of the variables were collected for every Ibis sampled, and the data still needs cleaning to accurately detail the sample size for each analysis.

Is there a difference in HaeParasit between Ibis of different ages? - Dependent variable (outcome): HaeParasit - Independent variable (predictor): Age

Is there a difference in HaeParasit between Ibis of different sex? - Dependent variable (outcome): HaeParasit - Independent variable (predictor): Sex

Is there a different in HaeParasit in Ibis sampled at different sites? - Dependent variable (outcome): HaeParasit - Independent variable (predictor): Site

Is there a difference in HaeParasit between Ibis sampled in different habitat types? - Dependent variable (outcome): HaeParasit - Independent variable (predictor): HabType

Is there a difference in HaeParasit between Ibis sampled at different times? - Dependent variable (outcome): HaeParasit - Independent variables (predictors): Date, Season

## 1.4 Expected Analyses

I think that I will have to conduct multiple ANOVA’s during the analysis process. Linear and logistic regression models apply here too for some variables.I would like to explore the possibility of using the location data for each sample to conduct a geographical analysis with a map as the output. I will be continuing to contact those involved with the study as well to see what analyses they want to be conducted and what questions they want answered.

## 1.5 Data Cleaning

The data has already been imported, so now we move onto the cleaning stage. The following code will clean, process, and save the raw Ibis blood datasets from 2010-2014 and 2015-2017. This section is contained in the “ProcessingScriptIbisBloodData.R” file.

##############################Cleaning the rdIbisBlood10\_14 data###########################  
IbisBlood10\_14 <- rename(rdIbisBlood10\_14, ID = X.) #renames the ID column  
  
#Extracts just the columns of interest  
IbisBlood10\_14 <- select(IbisBlood10\_14, ID, Site.ID, Date, PCR.sex, Age..adult.juvenile.,   
 Total.RBC, Haemoproteus.parasitemia.)  
  
#Renames the columns of interest  
IbisBlood10\_14 <- rename(IbisBlood10\_14, Site = Site.ID, Sex = PCR.sex, Age = Age..adult.juvenile.,  
 TotalRBC = Total.RBC, HaeParasit = Haemoproteus.parasitemia.)  
  
#Change "adult" and "juvenile" in the Age column to "A" and "J"  
IbisBlood10\_14$Age[IbisBlood10\_14$Age == "adult"] <- "A"  
IbisBlood10\_14$Age[IbisBlood10\_14$Age == "Adult"] <- "A"  
IbisBlood10\_14$Age[IbisBlood10\_14$Age == "juvenile"] <- "J"  
IbisBlood10\_14$Age[IbisBlood10\_14$Age == "SA"] <- "A"  
#Now the Ibises that had age recorded are referred to as "A" for adult or "J" for juvenile  
  
  
#Changing the DIV/0 values in HaeParasit to NA and change the column to numerical values  
IbisBlood10\_14$HaeParasit <- as.numeric(IbisBlood10\_14$HaeParasit) # Divide-by-zero errors are now NA

## Warning: NAs introduced by coercion

#Reformatting the date values in the Date column  
IbisBlood10\_14$Date <- excel\_numeric\_to\_date(IbisBlood10\_14$Date)  
#After this step, the values that just state the year are labelled with "1905", may need to go back or exclude from analysis  
  
#Abbreviating the site names to make analysis easier  
IbisBlood10\_14$Site[IbisBlood10\_14$Site == "Royal Palms"] <- "RP"  
IbisBlood10\_14$Site[IbisBlood10\_14$Site == "Everglades"] <- "E"  
IbisBlood10\_14$Site[IbisBlood10\_14$Site == "Dreher Park"] <- "DRP"  
IbisBlood10\_14$Site[IbisBlood10\_14$Site == "Indian Creek"] <- "ICP"  
IbisBlood10\_14$Site[IbisBlood10\_14$Site == "Indian Creek "] <- "ICP"  
IbisBlood10\_14$Site[IbisBlood10\_14$Site == "Indian Creek Park"] <- "ICP"  
IbisBlood10\_14$Site[IbisBlood10\_14$Site == "Juno Beach"] <- "JB"  
IbisBlood10\_14$Site[IbisBlood10\_14$Site == "Lion Country Safari"] <- "LCS"  
IbisBlood10\_14$Site[IbisBlood10\_14$Site == "Lake Worth"] <- "LW"  
IbisBlood10\_14$Site[IbisBlood10\_14$Site == "Prosperity Oaks"] <- "PO"  
IbisBlood10\_14$Site[IbisBlood10\_14$Site == "Promenade Plaza"] <- "PP"  
IbisBlood10\_14$Site[IbisBlood10\_14$Site == "San Mateo"] <- "SM"  
IbisBlood10\_14$Site[IbisBlood10\_14$Site == "Solid Waste Authority"] <- "SWA"  
IbisBlood10\_14$Site[IbisBlood10\_14$Site == "West Palm Beach Zoo"] <- "WPBZ"  
  
  
#Reformatting the mislabeled dates. These are set to Jan 1st of the year they were  
#recorded on the data sheet, these entries only had the year stated, not month and day.  
IbisBlood10\_14$Date[IbisBlood10\_14$Date == "1905-07-02"] <- "2010-01-01"  
IbisBlood10\_14$Date[IbisBlood10\_14$Date == "1905-07-03"] <- "2011-01-01"  
IbisBlood10\_14$Date[IbisBlood10\_14$Date == "1905-07-04"] <- "2012-01-01"  
  
#The bottom of the dataset has a couple rows of just missing data for all variables,  
#This code filters the dataset so that if any rows are present that has at least  
#one variable be NOT NA, R keeps it  
IbisBlood10\_14 <- filter(IbisBlood10\_14, if\_any(everything(), ~ !is.na(.)))  
  
#This completes the cleaning phase for IbisBlood10\_14  
#------------------------------------------------------------------------------------#  
  
  
############################Cleaning the rdIbisBlood15\_17 Data#########################  
  
IbisBlood15\_17<- rename(rdIbisBlood15\_17, ID= X.) #renames the ID column  
  
####Extracting only the columns relevant for analysis  
IbisBlood15\_17 <- select(IbisBlood15\_17,ID,Site.Name,Date,PCR.sex,Age,Season,Habitat.type,  
 Total.RBC,X..Haemoproteus,Haemoproteus.parasitemia.,F1,F2,F3,  
 Avg.RBC.FOV,Ibis..,Total.mass..g.,Mass.bag..g.,Mass.bird..g.,Body.condition.score..1.5.,  
 Ectoparasite.score..1.5.,Culmen.length..mm.,Wing.chord.length..mm.,  
 Tarsus.length..mm.,Tarsus.width..mm.)  
   
####Renaming the columns   
IbisBlood15\_17 <- rename(IbisBlood15\_17,Site = Site.Name,Sex = PCR.sex,HabType = Habitat.type,TotalRBC = Total.RBC,NumHae = X..Haemoproteus,  
 HaeParasit = Haemoproteus.parasitemia.,AvgRBC = Avg.RBC.FOV,  
 IbisNum = Ibis..,TotalMassG = Total.mass..g.,BagMassG = Mass.bag..g.,  
 BirdMassG = Mass.bird..g.,BodyCondScore = Body.condition.score..1.5.,  
 EctoParasitScore = Ectoparasite.score..1.5.,CulmenLmm = Culmen.length..mm.,  
 WingChordLmm = Wing.chord.length..mm.,TarsusLmm = Tarsus.length..mm.,  
 TarsusWmm = Tarsus.width..mm.)  
  
####Site column  
#Ibis that have sites matched with the 10\_14 dataset will have the same labels  
IbisBlood15\_17$Site[IbisBlood15\_17$Site == "Juno Beach"] <- "JB"  
IbisBlood15\_17$Site[IbisBlood15\_17$Site == "Indian Creek Park"] <- "ICP"  
IbisBlood15\_17$Site[IbisBlood15\_17$Site == "Dubois park"] <- "DUP"  
IbisBlood15\_17$Site[IbisBlood15\_17$Site == "Dreher Park"] <- "DRP"  
IbisBlood15\_17$Site[IbisBlood15\_17$Site == "Lion Country Safari"] <- "LCS"  
IbisBlood15\_17$Site[IbisBlood15\_17$Site == "Dubois Park"] <- "DUP"  
IbisBlood15\_17$Site[IbisBlood15\_17$Site == "Loxahatchee Wildlife Refuge"] <- "LOXWR"  
IbisBlood15\_17$Site[IbisBlood15\_17$Site == "J.W. Corbett Wildlife Management Area"] <- "JWC"  
IbisBlood15\_17$Site[IbisBlood15\_17$Site == "Gaines Park"] <- "GP"  
IbisBlood15\_17$Site[IbisBlood15\_17$Site == "Solid Waste Authority"] <- "SWA"  
IbisBlood15\_17$Site[IbisBlood15\_17$Site == "Fisheating creek"] <- "FC"  
IbisBlood15\_17$Site[IbisBlood15\_17$Site == "Kitching Creek"] <- "KC"  
IbisBlood15\_17$Site[IbisBlood15\_17$Site == "Loxahatchee NE"] <- "LOXNE"  
IbisBlood15\_17$Site[IbisBlood15\_17$Site == "DuBois Park"] <- "DUP"  
IbisBlood15\_17$Site[IbisBlood15\_17$Site == "TetraTech"] <- "TT"  
IbisBlood15\_17$Site[IbisBlood15\_17$Site == "Green Cay"] <- "GC"  
IbisBlood15\_17$Site[IbisBlood15\_17$Site == "Dreher park"] <- "DRP"  
IbisBlood15\_17$Site[IbisBlood15\_17$Site == "CROW Wildlife Rehabilitation Center"] <- "CROW"  
IbisBlood15\_17$Site[IbisBlood15\_17$Site == "Lake Worth"] <- "LW"  
IbisBlood15\_17$Site[IbisBlood15\_17$Site == "Royal Palm"] <- "RP"  
IbisBlood15\_17$Site[IbisBlood15\_17$Site == "JW Corbett"] <- "JWC"  
IbisBlood15\_17$Site[IbisBlood15\_17$Site == "Busch Wildlife Sanctuary"] <- "BWS"  
IbisBlood15\_17$Site[IbisBlood15\_17$Site == "Loxahatchee Slough"] <- "LOXS"  
IbisBlood15\_17$Site[IbisBlood15\_17$Site == "Indian Creek"] <- "ICP"  
IbisBlood15\_17$Site[IbisBlood15\_17$Site == "Cat House (Bonnie's, Richard Lane)"] <- "CAT"  
IbisBlood15\_17$Site[IbisBlood15\_17$Site == "Palm Beach Zoo"] <- "WPBZ"  
  
####Date column  
#Changing the date column to numeric values to allow date extraction  
IbisBlood15\_17$Date <- as.numeric(IbisBlood15\_17$Date)   
##Reformatting the date values in the Date column  
IbisBlood15\_17$Date <- excel\_numeric\_to\_date(IbisBlood15\_17$Date)  
  
####Sex column  
IbisBlood15\_17$Sex[IbisBlood15\_17$Sex == "not bled"] <- NA  
#Two values have missing data because the birds were not bled to determine sex  
IbisBlood15\_17$Sex[IbisBlood15\_17$Sex == "M?"] <- NA  
IbisBlood15\_17$Sex[IbisBlood15\_17$Sex == "F?"] <- NA  
IbisBlood15\_17$Sex[IbisBlood15\_17$Sex == "F "] <- "F"  
#Two values have question marks for unsure sex determination, removing those values  
  
####Age column  
#When an ibis reaches 4 years old, their plumage becomes all white. Any Ibis  
#that was sampled with brown plumage is considered a juvenile. Since Haemoproteus is  
#a chronic disease that mostly likely accumulates over time, long-lived birds should  
#theoretically have more parasitemia. Could not be true though and may come back to  
#look at the individual years in the juvenile birds  
IbisBlood15\_17$Age[IbisBlood15\_17$Age == "Juvenile"] <- "J"  
IbisBlood15\_17$Age[IbisBlood15\_17$Age == "Adult"] <- "A"  
IbisBlood15\_17$Age[IbisBlood15\_17$Age == "2nd/3rd year"] <- "J"  
IbisBlood15\_17$Age[IbisBlood15\_17$Age == "3rd year"] <- "J"  
IbisBlood15\_17$Age[IbisBlood15\_17$Age == "1st year"] <- "J"  
IbisBlood15\_17$Age[IbisBlood15\_17$Age == "adult"] <- "A"  
IbisBlood15\_17$Age[IbisBlood15\_17$Age == "2nd year"] <- "J"  
IbisBlood15\_17$Age[IbisBlood15\_17$Age == "NA"] <- NA  
IbisBlood15\_17$Age[IbisBlood15\_17$Age == "2nd Year"] <- "J"  
IbisBlood15\_17$Age[IbisBlood15\_17$Age == "Late third year"] <- "J"  
IbisBlood15\_17$Age[IbisBlood15\_17$Age == "ADULT"] <- "A"  
IbisBlood15\_17$Age[IbisBlood15\_17$Age == "Adult bc"] <- "A" #Not sure what bc stands for  
  
#Season and HabType are fine the way they are  
  
####TotalRBC  
IbisBlood15\_17$TotalRBC <- as.numeric(IbisBlood15\_17$TotalRBC)

## Warning: NAs introduced by coercion

#DIV/0 errors are now NA in this column  
  
#NumHae column is fine  
  
####HaeParasit  
IbisBlood15\_17$HaeParasit <- as.numeric(IbisBlood15\_17$HaeParasit)

## Warning: NAs introduced by coercion

####-999 values  
IbisBlood15\_17[IbisBlood15\_17 == -999] <- NA  
  
####Removing the last couple empty rows  
IbisBlood15\_17 <- IbisBlood15\_17[-c(401, 402, 403, 404, 405, 406),]  
  
#This completes the cleaning phase for IbisBlood15\_17  
#------------------------------------------------------------------------------------#  
  
###############################Saving the Cleaned Datasets###########################  
  
#Saving the Ibis Blood 10-14 dataset as an RDS in the processed data folder  
save\_data\_location <- here::here("processed\_data", "processedIbisBlood10\_14.rds")  
saveRDS(IbisBlood10\_14, file = save\_data\_location)  
  
#Saving the Ibis Blood 15-17 dataset as an RDS in the processed data folder  
save\_data\_location2 <- here::here("processed\_data", "processedIbisBlood15\_17.rds")  
saveRDS(IbisBlood15\_17, file = save\_data\_location2)

This next section will clean, process, and save the raw Ibis Field dataset. This section is contained in the “ProcessingScriptIbisFieldData.R” file.

######################### #Cleaning the rdIbisField Data###############################  
  
IbisField <- rename(rdIbisField, ID = ID..NSFSITE...) #renames the ID column  
  
#Selecting the columns of interest to this analysis from the datasheet  
IbisField <- select(IbisField,ID,Site.Name,Latitude,Longitude,Date,PCR.Sex,Age,Season,  
 Habitat.type,Ibis.number,Mass.bird..g.,Body.condition.score..1.5.,  
 Ectoparasite.score..1.5.,Culmen.length..mm.,Wing.chord.length..mm.,  
 Tarsus.length..mm.,Tarsus.width..mm.,Habituation.score.of.flock..1.5.)  
  
#Renaming the columns of interest, also matching the other cleaned datasets  
IbisField <- rename(IbisField,Site = Site.Name,Lat = Latitude,Long = Longitude,Sex = PCR.Sex,  
 HabType = Habitat.type,IbisNum = Ibis.number,BirdMassG = Mass.bird..g.,  
 BodyCondScore = Body.condition.score..1.5.,EctoParasitScore = Ectoparasite.score..1.5.,  
 CulmenLmm = Culmen.length..mm.,WingChordLmm = Wing.chord.length..mm.,  
 TarsusLmm = Tarsus.length..mm.,TarsusWmm = Tarsus.width..mm.,HabFlockScore = Habituation.score.of.flock..1.5.)  
  
#This dataset has "-999" values which mean they are intentionally left blank,  
#Birds with an ID number of "-999" represent sites that didn't catch any birds,  
#but had weather data collected. This code deletes those observations.  
IbisField <- IbisField[!(IbisField$ID == "-999"),]  
  
####Site  
#These site location abbreviations are the same across datasets  
IbisField$Site[IbisField$Site == "Busch Wildlife Sanctuary"] <- "BWS"  
IbisField$Site[IbisField$Site == "Cat House (Bonnie's, Richard Lane)"] <- "CAT"  
IbisField$Site[IbisField$Site == "Dubois Park"] <- "DUP"  
IbisField$Site[IbisField$Site == "DuBois Park"] <- "DUP"  
IbisField$Site[IbisField$Site == "Dreher Park"] <- "DRP"  
IbisField$Site[IbisField$Site == "Fisheating Creek"] <- "FC"  
IbisField$Site[IbisField$Site == "Green Cay"] <- "GC"  
IbisField$Site[IbisField$Site == "Gaines Park"] <- "GP"  
IbisField$Site[IbisField$Site == "Indian Creek"] <- "ICP"  
IbisField$Site[IbisField$Site == "Juno Beach"] <- "JB"  
IbisField$Site[IbisField$Site == "Juno Beach "] <- "JB"  
IbisField$Site[IbisField$Site == "J.W. Corbett Wildlife Management Area"] <- "JWC"  
IbisField$Site[IbisField$Site == "Kitching Creek"] <- "KC"  
IbisField$Site[IbisField$Site == "Lion Country Safari"] <- "LCS"  
IbisField$Site[IbisField$Site == "Lake Worth"] <- "LW"  
IbisField$Site[IbisField$Site == "Loxahatchee Slough"] <- "LOXS"  
IbisField$Site[IbisField$Site == "Loxahatchee Wildlife Refuge"] <- "LOXWR"  
IbisField$Site[IbisField$Site == "Loxahatchee/LILA Cell B2"] <- "LOXB2"  
IbisField$Site[IbisField$Site == "Loxahatchee NE "] <- "LOXNE"  
IbisField$Site[IbisField$Site == "Palm Beach Zoo"] <- "WPBZ"  
IbisField$Site[IbisField$Site == "Royal Palm"] <- "RP"  
IbisField$Site[IbisField$Site == "CROW Wildlife Rehabilitation Center"] <- "CROW"  
IbisField$Site[IbisField$Site == "Solid Waste Authority"] <- "SWA"  
IbisField$Site[IbisField$Site == "TetraTech"] <- "TT"  
  
####Sex Column  
#This replaces the -999 and FAIL values with NA. The IbisBlood datasheets carry more sex data for the birds  
IbisField$Sex[IbisField$Sex == "FAIL"] <- NA  
  
####Age Column  
IbisField$Age[IbisField$Age == "Adult"] <- "A"  
IbisField$Age[IbisField$Age == "3rd year"] <- "J"  
IbisField$Age[IbisField$Age == "2nd year"] <- "J"  
IbisField$Age[IbisField$Age == "1st year"] <- "J"  
IbisField$Age[IbisField$Age == "NA"] <- NA  
IbisField$Age[IbisField$Age == "Adult "] <- "A"  
IbisField$Age[IbisField$Age == "2nd year "] <- "J"  
IbisField$Age[IbisField$Age == "Juvenile"] <- "J"  
IbisField$Age[IbisField$Age == "2nd Year"] <- "J"  
  
#Replacing all -999 values in the dataset with NA  
IbisField[IbisField == -999] <- NA  
  
#Replacing NA text strings with regular NA values in multiple columns  
IbisField$BodyCondScore[IbisField$BodyCondScore == "NA"] <- NA  
IbisField$HabFlockScore[IbisField$HabFlockScore == "NA"] <- NA  
IbisField$CulmenLmm[IbisField$CulmenLmm == "NA"] <- NA  
IbisField$WingChordLmm[IbisField$WingChordLmm == "NA"] <- NA  
IbisField$TarsusLmm[IbisField$TarsusLmm == "NA"] <- NA  
IbisField$TarsusWmm[IbisField$TarsusWmm == "NA"] <- NA  
  
#--------------------------------------------------------------------------------------#  
  
###############################Saving the Cleaned Dataset###############################  
  
save\_data\_location3 <- here::here("processed\_data", "processedIbisFielddata.rds")  
saveRDS(IbisField, file = save\_data\_location3)

This cleaning code section will clean, process, and save the Ibis urbanization dataset. This section is contained in the “ProcessingScriptIbisUrban.R” file.

##############################Cleaning the rdUrbanIbis data###########################  
#Selects only the columns of interest in the dataset  
IbisUrban <- select(rdUrbanIbis, Name:Serotype)  
  
#Renames the columns of interest  
IbisUrban <- rename(IbisUrban,ID = Name,Date = Collection.Date,HgPPM = mg.kg.Hg..PPM.,  
 Site = Site.Name,UrbanPercent = Site...Urbanized)  
  
#Change the site names to their abbreviations  
IbisUrban$Site[IbisUrban$Site == "Juno Beach"] <- "JB"  
IbisUrban$Site[IbisUrban$Site == "Indian Creek"] <- "ICP"  
IbisUrban$Site[IbisUrban$Site == "Dubois Park"] <- "DUP"  
IbisUrban$Site[IbisUrban$Site == "Dreher Park"] <- "DRP"  
IbisUrban$Site[IbisUrban$Site == "Lion Country Safari"] <- "LCS"  
IbisUrban$Site[IbisUrban$Site == "Loxahatchee Wildlife Refuge"] <- "LOXWR"  
IbisUrban$Site[IbisUrban$Site == "Solid Waste Authority "] <- "SWA"  
IbisUrban$Site[IbisUrban$Site == "Gaines Park"] <- "GP"  
IbisUrban$Site[IbisUrban$Site == "Kitching Creek"] <- "KC"  
IbisUrban$Site[IbisUrban$Site == "Kitching Creek "] <- "KC"  
IbisUrban$Site[IbisUrban$Site == "Loxahatchee NE"] <- "LOXNE"  
IbisUrban$Site[IbisUrban$Site == "TetraTech"] <- "TT"  
IbisUrban$Site[IbisUrban$Site == "Loxahatchee NE "] <- "LOXNE"  
IbisUrban$Site[IbisUrban$Site == "Green Cay"] <- "GC"  
IbisUrban$Site[IbisUrban$Site == "J.W. Corbett Wildlife Management Area"] <- "JWC"  
IbisUrban$Site[IbisUrban$Site == "Loxahatchee Slough"] <- "LOXS"  
IbisUrban$Site[IbisUrban$Site == "Lake Worth"] <- "LW"  
  
#--------------------------------------------------------------------------------------#  
  
################################Saving the UrbanIbis Data#############################  
#Saving the Ibis Blood 15-17 dataset as an RDS in the processed data folder  
save\_data\_location3 <- here::here("processed\_data", "processedIbisUrban.rds")  
saveRDS(IbisUrban, file = save\_data\_location3)  
  
#--------------------------------------------------------------------------------------#

This last cleaning section combines the data sheets into one to use for analysis. The urbanization gradient has not been created as of yet, but a meeting this month with those involved with the project will help shed some light on the exact specifics for analysis. This code is located in the “ExploratoryAnalysisScript.R” file.

###############################Completing Combination Data Frames#########################  
#Outer joining IbisBlood10\_14 and IbisBlood15\_17, keeping all values of the 10\_14 dataset  
IbisBlood10\_17 <- merge(IbisBlood10\_14, IbisBlood15\_17, by = c("ID", "Site", "Date", "Sex", "Age",  
 "TotalRBC", "HaeParasit"), all=TRUE)  
#There is no overlap in the ID numbers between the two, so no recapture is assumed.  
  
#This is a left join, this data frame below only contains birds that were in the IbisBlood10\_17  
#dataset (all with parasitemia data), with additional information added by the IbisField data. Ibis  
#from the Field data that weren't tested for parasitemia or weren't included on Blood dataset  
#are not included in this dataframe  
  
#IbisBlood10\_17(with Field)\_(Parasitemia data)  
IbisC <- left\_join(IbisBlood10\_17, IbisField, by = c("ID", "Site", "Date", "Sex", "Age", "Season", "HabType",  
 "IbisNum", "BirdMassG", "BodyCondScore", "EctoParasitScore",  
 "CulmenLmm", "WingChordLmm", "TarsusLmm", "TarsusWmm"))  
  
#This section classifies the habitat type for the sites without a habitat type.  
IbisC <- IbisC %>%  
 mutate(HabType = ifelse(Site == "ICP", "Urban", HabType)) %>%  
 mutate(HabType = ifelse(Site == "JB", "Urban", HabType)) %>%  
 mutate(HabType = ifelse(Site == "RP", "Urban", HabType)) %>%  
 mutate(HabType = ifelse(Site == "LCS", "Urban", HabType)) %>%  
 mutate(HabType = ifelse(Site == "LW", "Urban", HabType)) %>%  
 mutate(HabType = ifelse(Site == "SWA", "Urban", HabType)) %>%  
 mutate(HabType = ifelse(Site == "E", "Natural", HabType)) %>%  
 mutate(HabType = ifelse(Site == "WPBZ", "Urban", HabType)) %>%  
 mutate(HabType = ifelse(Site == "PO", "Urban", HabType)) %>%  
 mutate(HabType = ifelse(Site == "PP", "Urban", HabType)) %>%  
 mutate(HabType = ifelse(Site == "DRP", "Urban", HabType)) %>%  
 mutate(HabType = ifelse(Site == "SM", "Urban", HabType))  
   
   
##This section is adding in the seasons for observations that have collection dates but no season.  
  
#Creating intervals for adding seasons based on the date column  
int1 <- interval(ymd("2012-12-01"), ymd("2012-12-31"))  
int2 <- interval(ymd("2013-06-01"), ymd("2013-08-31"))  
int3 <- interval(ymd("2013-12-01"), ymd("2013-12-31"))  
int4 <- interval(ymd("2014-02-01"), ymd("2014-03-31"))  
  
#Adding seasons to the ibis without them stated based on the date the sameple was collected  
IbisC <- IbisC %>%  
 mutate(Season = ifelse(Date %within% int1, "Fall 2012", Season)) %>%  
 mutate(Season = ifelse(Date %within% int2, "Summer 2013", Season)) %>%  
 mutate(Season = ifelse(Date %within% int3, "Fall 2013", Season)) %>%  
 mutate(Season = ifelse(Date %within% int4, "Spring 2014", Season))  
  
  
#Adding a column for HaeParasit Presence/Absence data  
IbisC <- IbisC %>%  
 mutate(HaeParasitPA = as.integer(HaeParasit > 0 & !is.na(HaeParasit)))  
  
#Adding a column for transformed HaeParasit data on a logarithm base 10 scale,  
#then removing the -Inf values due to there being no parasitemia in those birds. This column  
#then only contains logarithm transformed parasitemia data for birds with a positive value  
#of parasitemia.  
IbisC <- IbisC %>%  
 mutate(HaeParasitLog10 = log10(HaeParasit)) %>%  
 mutate(HaeParasitLog10 = ifelse(HaeParasitLog10 == "-Inf", NA, HaeParasitLog10))  
  
#Changing the BodyCondScore to an ordinal scale  
IbisC <- IbisC %>%  
 mutate(BodyCondScore = recode\_factor(BodyCondScore, '1' = 'Emaciated',  
 '2' = 'Underweight', '3' = 'Ideal',  
 '4' = 'Overweight', '5' = 'Obese',  
 .ordered = TRUE))  
  
#Changing the class of some variables for analysis. From character to numeric.  
C2N <- c(16,17,18,21,22,23,24)  
  
IbisC[ , C2N] <- apply(IbisC[ , C2N], 2,  
 function(x) as.numeric(as.character(x)))

## Warning in FUN(newX[, i], ...): NAs introduced by coercion  
  
## Warning in FUN(newX[, i], ...): NAs introduced by coercion  
  
## Warning in FUN(newX[, i], ...): NAs introduced by coercion  
  
## Warning in FUN(newX[, i], ...): NAs introduced by coercion  
  
## Warning in FUN(newX[, i], ...): NAs introduced by coercion  
  
## Warning in FUN(newX[, i], ...): NAs introduced by coercion  
  
## Warning in FUN(newX[, i], ...): NAs introduced by coercion

This concludes the data cleaning section, we are now ready for the exploratory analysis!

## 1.6 Exploratory analysis

### 1.6.1 Tables

To start the exploratory analysis, we’re going to create some quick tables to get an idea of the metrics for some of the columns of interest.

IbisC %>%  
 count(Sex, wt = HaeParasitPA)

## Sex n  
## 1 F 187  
## 2 M 96  
## 3 <NA> 78

Looking at presence/absence data for haemoproteus parasitemia, there were about double the number of female ibises with active infection compared to the males. 78 of the ibises sampled were negative for haemoproteus infection.

IbisC %>%  
 count(Age, wt = HaeParasitPA)

## Age n  
## 1 A 220  
## 2 J 82  
## 3 <NA> 59

Here we can see that there was about 2.7 times more adults sampled with active infection than juvenile ibises. 59 of the sampled ibises with active infection either couldn’t be aged or age was not recorded.

IbisC %>%  
 count(Age, Sex, wt = HaeParasitPA)

## Age Sex n  
## 1 A F 109  
## 2 A M 51  
## 3 A <NA> 60  
## 4 J F 48  
## 5 J M 32  
## 6 J <NA> 2  
## 7 <NA> F 30  
## 8 <NA> M 13  
## 9 <NA> <NA> 16

Combining age and sex, we can see that adult females had the highest number of active infections. Categories with missing numbers are prevalent.

IbisC %>%  
 count(Site, wt = HaeParasitPA)

## Site n  
## 1 BWS 1  
## 2 CAT 3  
## 3 CROW 1  
## 4 DRP 40  
## 5 DUP 10  
## 6 E 0  
## 7 FC 0  
## 8 GC 27  
## 9 GP 20  
## 10 ICP 48  
## 11 JB 31  
## 12 JWC 5  
## 13 KC 8  
## 14 LCS 34  
## 15 LOXNE 11  
## 16 LOXS 8  
## 17 LOXWR 6  
## 18 LW 15  
## 19 PO 17  
## 20 PP 7  
## 21 RP 0  
## 22 SM 1  
## 23 SWA 59  
## 24 TT 7  
## 25 WPBZ 0  
## 26 <NA> 2

Looking at site, SWA, ICP, and DRP were the sites with the highest number of sampled ibises with active infection.

IbisC %>%  
 count(Season, wt = HaeParasitPA)

## Season n  
## 1 Fall 2012 16  
## 2 Fall 2013 15  
## 3 Fall 2015 35  
## 4 Fall 2016 39  
## 5 Spring 2014 40  
## 6 Spring 2016 65  
## 7 Spring 2017 41  
## 8 Summer 2013 44  
## 9 Summer 2016 35  
## 10 Summer 2017 15  
## 11 <NA> 16

Looking at season sampled, Spring 2016 had the highest number of sampled ibises with active infection, followed by Summer 2013, Spring 2017, and Spring 2014.

IbisC %>%  
 count(BodyCondScore, Age, wt = HaeParasitPA)

## BodyCondScore Age n  
## 1 Emaciated A 1  
## 2 Emaciated J 2  
## 3 Underweight A 29  
## 4 Underweight J 34  
## 5 Ideal A 89  
## 6 Ideal J 20  
## 7 Overweight A 22  
## 8 Overweight J 14  
## 9 Overweight <NA> 1  
## 10 Obese A 1  
## 11 2.5 A 1  
## 12 NA A 0  
## 13 1.5 J 1  
## 14 <NA> A 77  
## 15 <NA> J 11  
## 16 <NA> <NA> 58

Looking at body condition score, the majority of the birds sampled with active infection were an ideal weight or underweight. This lets us know that an active infection didn’t affect body condition too harshly on these birds.

### 1.6.2 Figures

Now we can move onto creating some figures for the data to see any patterns visually.

Figure 1: Haemoproteus parasitemia from sampled infected Ibises from 2010-2017 by habitat type

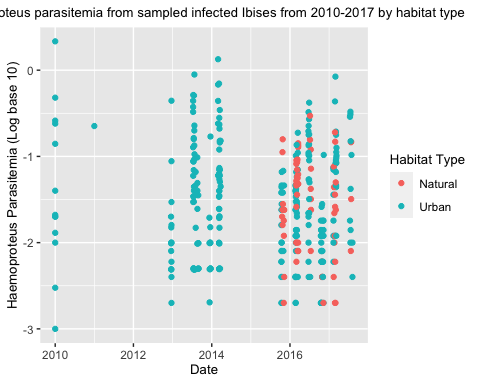
ExpFigure1 <-ggplot(IbisC) +  
 geom\_point(aes(x = Date, y = HaeParasitLog10, color = HabType)) +  
 ggtitle("Haemoproteus parasitemia from sampled infected Ibises from 2010-2017 by habitat type") +  
 xlab("Date") +  
 ylab("Haemoproteus Parasitemia (Log base 10)") +  
 scale\_color\_discrete("Habitat Type", na.translate = F) +  
 theme(plot.title = element\_text(hjust = 0.5, size = 10),  
 axis.title.x = element\_text(size = 10),  
 axis.title.y = element\_text(size = 10),  
 legend.title = element\_text(size = 10))  
  
Figure\_file1 <- here("results", "Figures", "ExpFigure1.png")  
ggsave(filename = Figure\_file1,plot = ExpFigure1)

## Saving 5 x 4 in image

## Warning: Removed 309 rows containing missing values (geom\_point).

ExpFigure1

## Warning: Removed 309 rows containing missing values (geom\_point).



This scatterplot graphs the sample dates chronologically against the parasitemia measured from the sampled ibis on a transformed log base 10 scale. Habitat type is depicted by color of the dots. Overall, there is a general narrowing of the data over time with a slight negative correlation. There isn’t a clear difference between natural and urban habitat types when it comes to amount of parasitemia.

Figure 2: Haemoproteus parasitemia from adult and juvenile sampled Ibises

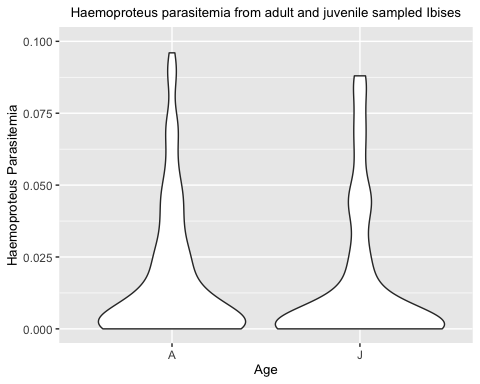
ExpFigure2 <- IbisC %>%  
 drop\_na(Age) %>%  
 ggplot() +  
 geom\_violin(aes(x = Age, y = HaeParasit)) +  
 ggtitle("Haemoproteus parasitemia from adult and juvenile sampled Ibises") +  
 xlab("Age") +  
 ylab("Haemoproteus Parasitemia") +  
 scale\_y\_continuous(limits = c(0, 0.1)) +  
 theme(plot.title = element\_text(hjust = 0.5, size = 10),  
 axis.title.x = element\_text(size = 10),  
 axis.title.y = element\_text(size = 10),  
 legend.title = element\_text(size = 10))  
  
Figure\_file2 <- here("results", "Figures", "ExpFigure2.png")  
ggsave(filename = Figure\_file2, plot = ExpFigure2)

## Saving 5 x 4 in image

## Warning: Removed 142 rows containing non-finite values (stat\_ydensity).

ExpFigure2

## Warning: Removed 142 rows containing non-finite values (stat\_ydensity).



This violin plot graphs the haemoproteus parasitemia distribution for the adult and juvenile Ibises sampled. The majority of the parasitemia values are on the lower end for both adults and juveniles. The distribution for adults reaches higher than juveniles. The overall structure for the adult distribution is higher and skinnier, while juveniles have more of their parasitemias lower in value.

Figure 3: Haemoproteus parasitemia from sampled Ibises by sex

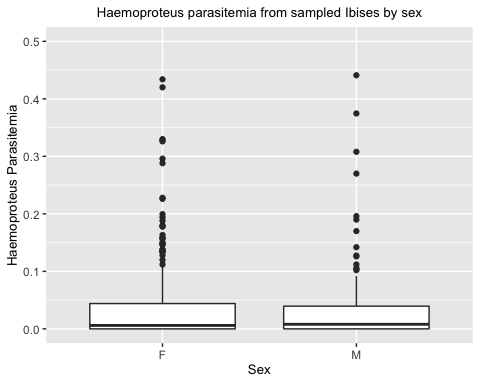
ExpFigure3 <- IbisC %>%  
 drop\_na(Sex) %>%  
 ggplot() +  
 geom\_boxplot(aes(x = Sex, y = HaeParasit)) +  
 scale\_y\_continuous(limits = c(0, 0.5)) +  
 ggtitle("Haemoproteus parasitemia from sampled Ibises by sex") +  
 xlab("Sex") +  
 ylab("Haemoproteus Parasitemia") +  
 theme(plot.title = element\_text(hjust = 0.5, size = 10),  
 axis.title.x = element\_text(size = 10),  
 axis.title.y = element\_text(size = 10),  
 legend.title = element\_text(size = 10))  
  
Figure\_file3 <- here("results", "Figures", "ExpFigure3.png")  
ggsave(filename = Figure\_file3, plot = ExpFigure3)

## Saving 5 x 4 in image

## Warning: Removed 21 rows containing non-finite values (stat\_boxplot).

ExpFigure3

## Warning: Removed 21 rows containing non-finite values (stat\_boxplot).



The haemoproteus parasitemia distributions between sexes is extremely similar. The median value for males is slightly higher than females. The majority of the measured parasitemias are between 0.0 and 0.1.

Figure 4: Haemoproteus parasitemia by total female and male Ibis mass

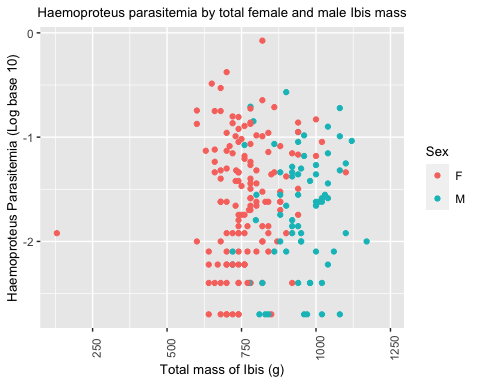
ExpFigure4 <- IbisC %>%  
 drop\_na(BirdMassG, Sex) %>%  
 ggplot() +  
 geom\_point(aes(x = BirdMassG, y = HaeParasitLog10, color = Sex)) +  
 theme(axis.text.x = element\_text(angle = 90)) +  
 scale\_x\_continuous(limits = c(130, 1240)) +  
 ggtitle("Haemoproteus parasitemia by total female and male Ibis mass") +  
 xlab("Total mass of Ibis (g)") +  
 ylab("Haemoproteus Parasitemia (Log base 10)") +  
 theme(plot.title = element\_text(hjust = 0.5, size = 10),  
 axis.title.x = element\_text(size = 10),  
 axis.title.y = element\_text(size = 10),  
 legend.title = element\_text(size = 10))  
  
Figure\_file4 <- here("results", "Figures", "ExpFigure4.png")  
ggsave(filename = Figure\_file4, plot = ExpFigure4)

## Saving 5 x 4 in image

## Warning: Removed 107 rows containing missing values (geom\_point).

ExpFigure4

## Warning: Removed 107 rows containing missing values (geom\_point).



Looking at the plot, male birds seem to weigh more than females. However, when comparing the parasitemia between females and males, there isn’t a clear difference between the two.

Figure 5: Haemoproteus parasitemia by female and male Ibis mass and body condition score

ExpFigure5 <- IbisC %>%  
 drop\_na(BirdMassG, BodyCondScore, Sex) %>%  
 ggplot() +  
 geom\_point(aes(x = BirdMassG, y = HaeParasitLog10, color = BodyCondScore)) +  
 facet\_wrap(~ Sex) +  
 ggtitle("Haemoproteus parasitemia by female and male Ibis mass and body condition score") +  
 xlab("Total mass of Ibis (g)") +  
 ylab("Haemoproteus Parasitemia (Log base 10)") +  
 theme(axis.text.x = element\_text(angle = 90),  
 plot.title = element\_text(hjust = 0.5, size = 10),  
 axis.title.x = element\_text(size = 10),  
 axis.title.y = element\_text(size = 10),  
 legend.title = element\_text(size = 10)) +  
 scale\_x\_continuous(limits = c(130, 1240)) +  
 scale\_color\_discrete("Body Condition", na.translate = F)  
  
Figure\_file5 <- here("results", "Figures", "ExpFigure5.png")  
ggsave(filename = Figure\_file5, plot = ExpFigure5)

## Saving 5 x 4 in image

## Warning: Removed 107 rows containing missing values (geom\_point).

ExpFigure5

## Warning: Removed 107 rows containing missing values (geom\_point).



Looking at this plot, there are more underweight female ibises with higher parasitemias compared to males. This could be due to the higher number of sampled females though.

Figure 6: Haemoproteus parasitemia by sample site and habitat type

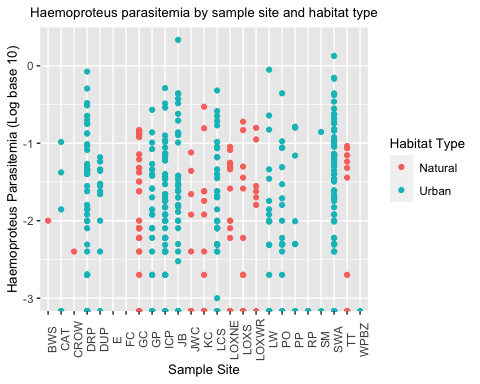
ExpFigure6 <- IbisC %>%  
 drop\_na(HabType, Site) %>%  
 ggplot() +  
 geom\_point(aes(x = Site, y = log10(HaeParasit), color = HabType)) +  
 theme(axis.text.x = element\_text(angle = 90),  
 plot.title = element\_text(hjust = 0.5, size = 10),  
 axis.title.x = element\_text(size = 10),  
 axis.title.y = element\_text(size = 10),  
 legend.title = element\_text(size = 10)) +  
 ggtitle("Haemoproteus parasitemia by sample site and habitat type") +  
 xlab("Sample Site") +  
 ylab("Haemoproteus Parasitemia (Log base 10)") +  
 scale\_color\_discrete("Habitat Type", na.translate = F)   
  
Figure\_file6 <- here("results", "Figures", "ExpFigure6.png")  
ggsave(filename = Figure\_file6, plot = ExpFigure6)

## Saving 5 x 4 in image

## Warning: Removed 121 rows containing missing values (geom\_point).

ExpFigure6

## Warning: Removed 121 rows containing missing values (geom\_point).



Looking at this plot broadly, it seems that the sites with a natural habitat have a slightly lower average parasitemia than the sites with an urban habitat.

Overall, there are no huge patterns or clear differences between covariates and variates so far. Statistical analyses may pick up on any significant differences later on, but I will return to the exploratory analysis as I continue with the project. I will also be meeting with those involved with the project to discuss as well.

## 1.7 Statistical Analyses

### 1.7.1 Presence/Absence Haemoproteus Parasitemia Data

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. I am currently struggling with applying the models generated by CV to the testing data.

data\_locationPA1 <- here::here("results", "Tables", "Model1MetricsPA.Rda")  
table1 <- readRDS(data\_locationPA1)  
table1

## # A tibble: 2 × 6  
## .metric .estimator mean n std\_err .config   
## <chr> <chr> <dbl> <int> <dbl> <chr>   
## 1 accuracy binary 0.535 10 0.0158 Preprocessor1\_Model1  
## 2 roc\_auc binary 0.472 10 0.0161 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.

data\_locationPA2 <- here::here("results", "Tables", "Model2MetricsPA.Rda")  
table2 <- readRDS(data\_locationPA2)  
table2

## # A tibble: 2 × 6  
## .metric .estimator mean n std\_err .config   
## <chr> <chr> <dbl> <int> <dbl> <chr>   
## 1 accuracy binary 0.605 10 0.0217 Preprocessor1\_Model1  
## 2 roc\_auc binary 0.529 10 0.0375 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.

data\_locationPA3 <- here::here("results", "Tables", "Model3MetricsPA.Rda")  
table3 <- readRDS(data\_locationPA3)  
table3

## # A tibble: 2 × 6  
## .metric .estimator mean n std\_err .config   
## <chr> <chr> <dbl> <int> <dbl> <chr>   
## 1 accuracy binary 0.569 10 0.0175 Preprocessor1\_Model1  
## 2 roc\_auc binary 0.453 10 0.0158 Preprocessor1\_Model1

This univariate classification GLM looked at the effect of age of the Ibis on the presence of Haemoproteus in infected birds. The results were unsatisfactory.

data\_locationPA4 <- here::here("results", "Tables", "Model4MetricsPA.Rda")  
table4 <- readRDS(data\_locationPA4)  
table4

## # A tibble: 2 × 6  
## .metric .estimator mean n std\_err .config   
## <chr> <chr> <dbl> <int> <dbl> <chr>   
## 1 accuracy binary 0.578 8 0.0172 Preprocessor1\_Model1  
## 2 roc\_auc binary 0.584 8 0.0177 Preprocessor1\_Model1

This univariate classification GLM looked at the effect of sampling site on the presence of Haemoproteus in infected birds. The results were unsatisfactory.

data\_locationPA5 <- here::here("results", "Tables", "Model5MetricsPA.Rda")  
table5 <- readRDS(data\_locationPA5)  
table5

## # A tibble: 2 × 6  
## .metric .estimator mean n std\_err .config   
## <chr> <chr> <dbl> <int> <dbl> <chr>   
## 1 accuracy binary 0.541 10 0.0181 Preprocessor1\_Model1  
## 2 roc\_auc binary 0.544 10 0.0180 Preprocessor1\_Model1

This univariate classification GLM looked at the effect of sampling date on the presence of Haemoproteus in infected birds. The results were unsatisfactory.

data\_locationPA6 <- here::here("results", "Tables", "Model6MetricsPA.Rda")  
table6 <- readRDS(data\_locationPA6)  
table6

## # A tibble: 2 × 6  
## .metric .estimator mean n std\_err .config   
## <chr> <chr> <dbl> <int> <dbl> <chr>   
## 1 accuracy binary 0.634 10 0.0331 Preprocessor1\_Model1  
## 2 roc\_auc binary 0.649 10 0.0360 Preprocessor1\_Model1

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

data\_locationPA7 <- here::here("results", "Tables", "Model7MetricsPA.Rda")  
table7 <- readRDS(data\_locationPA7)  
table7

## # A tibble: 2 × 6  
## .metric .estimator mean n std\_err .config   
## <chr> <chr> <dbl> <int> <dbl> <chr>   
## 1 accuracy binary 0.615 10 0.0263 Preprocessor1\_Model1  
## 2 roc\_auc binary 0.576 10 0.0424 Preprocessor1\_Model1

This multivariate classification GLM looked at the effect of age and sex on the presence of Haemoproteus in infected birds. The results were unsatisfactory, this model scored about the same as the previous univariate models for age and sex.

data\_locationPA8 <- here::here("results", "Tables", "Model8MetricsPA.Rda")  
table8 <- readRDS(data\_locationPA8)  
table8

## # A tibble: 2 × 6  
## .metric .estimator mean n std\_err .config   
## <chr> <chr> <dbl> <int> <dbl> <chr>   
## 1 accuracy binary 0.614 10 0.0299 Preprocessor1\_Model1  
## 2 roc\_auc binary 0.671 10 0.0387 Preprocessor1\_Model1

This multivariate classification GLM looked at the effect of sampling date and sampling season on the presence of Haemoproteus in infected birds. The results here had the highest performance of all the other models so far. One way to interpret this model would be that as the sampling effort continued from 2010-2017, the passage of time and season had a significant effect on the presence of Haemoproteus in Ibis. However, the performance of this model isn’t high enough to successfully conclude with that explanation.

### 1.7.2 Regression Models for the Amount of Haemoproteus Parasitemia

This section details the results of a preliminary statistical analysis on the Haemoproteus parasitemia data looking at the amount of measured parasitemia per sampled bird (on a logarithm base 10 scale), as opposed to presence/absence data.

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. I am currently struggling with applying the models generated by CV to the testing data.

#### 1.7.2.1 Random Forest Model

The following are the results from a random forest model run on all the predictors against the Haemoproteus parasitemia (log base 10) data. The missing data values were imputed by K nearest neighbors. The model was run on all 10 folds from the training dataset.

data\_locationRF <- here::here("results", "Tables", "RandomForestMetrics.Rda")  
table9 <- readRDS(data\_locationRF)  
table9

## # A tibble: 2 × 6  
## .metric .estimator mean n std\_err .config   
## <chr> <chr> <dbl> <int> <dbl> <chr>   
## 1 rmse standard 0.575 10 0.0211 Preprocessor1\_Model1  
## 2 rsq standard 0.218 10 0.0314 Preprocessor1\_Model1

This random forest model looked at the effect of all predictors on the amount of Haemoproteus in infected birds. The results were unsatisfactory, this model had a relatively high RMSE and low R squared value.

#### 1.7.2.2 Linear Regression Model

The following are the results from a linear regression model run on habitat type against the Haemoproteus parasitemia (log base 10) data.

data\_locationLM1 <- here::here("results", "Tables", "LinMetrics1.Rda")  
table10 <- readRDS(data\_locationLM1)  
table10

## # A tibble: 2 × 6  
## .metric .estimator mean n std\_err .config   
## <chr> <chr> <dbl> <int> <dbl> <chr>   
## 1 rmse standard 0.649 10 0.0255 Preprocessor1\_Model1  
## 2 rsq standard 0.0327 10 0.0126 Preprocessor1\_Model1

This linear regression model looked at the effect of habitat type on the amount of Haemoproteus in infected birds. The results were unsatisfactory, the RMSE is high and the R squared value is very low.

#### 1.7.2.3 Conclusions so far

This concludes the statistical analysis for part 3. Overall, the models did not perform well, especially those looking at the amount of Haemoproteus in infected birds. The presence/absence data has more potential for conclusions to be made based on generated models. The next steps would be to fine-tune the classification models for the presence/absence data and create a random forest model.

Current problems that have arose in this process include imputing missing values to run models (as there is a large amount of missing data in this dataset), and visualizing the tidymodels results on a graph.