Analyzing Animal Behaviors from their physical qualities

Jack Wang

11/18/2021

PREFACE

Mathematicians have long believed that nature can be explained with math, from the Fibonacci Sequence to hexagons. Ecologists collected several data points pertaining to the animal specimens that they caught over over a long period of time in collaboration with other ecologists. Using the data collect, relationships and patterns such as the hind foot length of animals of different latitudes, mating seasons, biological triggers for mating, and migratory patterns can be explored.

LIBRARY IMPORTING

Libraries that are included in our final project include: car, tdyverse, ggplot2, lubridate, tidymodels, knitr, readxl, ozmaps, sf, and caret.

DATA IMPORTING

Data importing was done in a standard manner. All data sets included in the original folder were imported into our program and modified as needed.

DATA CARPENTRY - OVERALL

1) Merges (dy > day, hfl > hindfoot_length, wgt > weight, mo > month, yr > year)
2) Deleting cols (note1, note2, note3, note4, note5, nestdir, neststk, prevlet, prevrt, ltag, tag, JSON, stake, plot, plot_id2) because they are not needed in any of the analysis that we will be performing.

HYPOTHESIS I

DATA CARPENTRY

Within the data, we only wanted HFL, age, species, sex, wgt, latitude, and longitude data. Thus the variables in question was selected out. Any data points with an NA in those category was dropped. This is because if there is no entry, we cannot use it to model. There was a species who's age was labeled as 'ZJ' where Z mean adult, and J mean Juvenile. Since both was select, we cannot decide with a good concious whether the animal is a juvenile or adult so it was dropped from the study.

A-PRORI HYPOTHESIS

Animals have longer hind foot lengths the closer they are to the equator.

PREFACE

We want to evaluate this claim as it is well known that animals in warmer regions are more active. The increased active necessitates the need for speed. Animals need more speed to catch prey and prey needs speed to escape predator.

Statistical Hypothesis:

H0 = No statistical significant difference in HFL Averages across different ages, species, sex, and distance from equator

Ha = At least one significant difference in HFL averages across different ages, species, sex, and distance from equator

The hypothesis will be tested based on the following equation:

$$\begin{aligned} HFL =& \beta_0 + \beta_1(AH?) + \beta_2(AS?) + \beta_3(BA?) + \beta_4(CB?) + \beta_5(CM?) + \\ & \beta_6(CQ?) + \beta_7(CS?) + \beta_8(CT?) + \beta_9(CU?) + \beta_{10}(CV?) + \\ & \beta_{11}(DM?) + \beta_{12}(DO?) + \beta_{13}(DS?) + \beta_{14}(DX?) + \beta_{15}(EO?) + \beta_{16}(GS?) + \beta_{17}(NL?) + \beta_{18}(NX?) + \\ & \beta_{19}(OL?) + \beta_{20}(OT?) + \beta_{21}(OX?) + \beta_{22}(PB?) + \beta_{23}(PC?) + \beta_{24}(PE?) + \beta_{25}(PF?) + \beta_{26}(PG?) + \\ & \beta_{27}(PH?) + \beta_{28}(PI?) + \beta_{29}(PL?) + \beta_{30}(PM?) + \beta_{31}(PP?) + \beta_{32}(PU?) + \beta_{33}(PX?) + \beta_{34}(RF?) + \\ & \beta_{35}(RM?) + \beta_{36}(RO?) + \beta_{37}(RX?) + \beta_{38}(SA?) + \beta_{39}(SB?) + \beta_{40}(SC?) + \beta_{41}(SF?) + \beta_{42}(SH?) + \\ & \beta_{43}(SO?) + \beta_{44}(SS?) + \beta_{45}(ST?) + \beta_{46}(SU?) + \beta_{47}(SX?) + \beta_{48}(UL?) + \beta_{49}(UP?) + \beta_{50}(UR?) + \\ & \beta_{51}(US?) + \beta_{52}(ZL?) + \beta_{53}(ZM?) + \beta_{54}(Longitude) + \\ & \beta_{55}(Latitude) + \beta_{56}(Male?) + \beta_{57}(Juvenile?) \end{aligned}$$

Scientific Hypothesis: Animals have longer hind foot lengths the closer they are to the equator

Apriori Hypothesis Evaluation

From the data set, if one was to filter out the 'NA's from the data set for each of the variables used in the model. One will find that there is not one data entry where there there is a complete data set. One of the biggest problems with this data set is the lack of latitude and longitude data entries. The entries are plotted on a U.S map and have been inserted into the supplementary to show that it lacks data points and that there are only located in Florida and Idaho. For that it is impossible to model the data set with the equation given early. Since this is the case, in our Exploratory Hypothesis, we will pivot our study from examining the effect of distance from equator on hind foot length to the weight of an animal on hind foot length and whether this relation is sustained even within a species.

EXPLORATORY HYPOTHESIS

Does the size of an animal effect the their hind foot length? In this exploration we will examine the general relationship between an animal's weight and hind foot length. Then, we will explore whether this relationship holds true within the species.

Scientific Hypothesis: The heavier the animal the longer their hind foot length. This is because the Hind foot is used to support the body. The bigger the animal the stronger the hind foot needs to be.

Statistical Hypothesis:

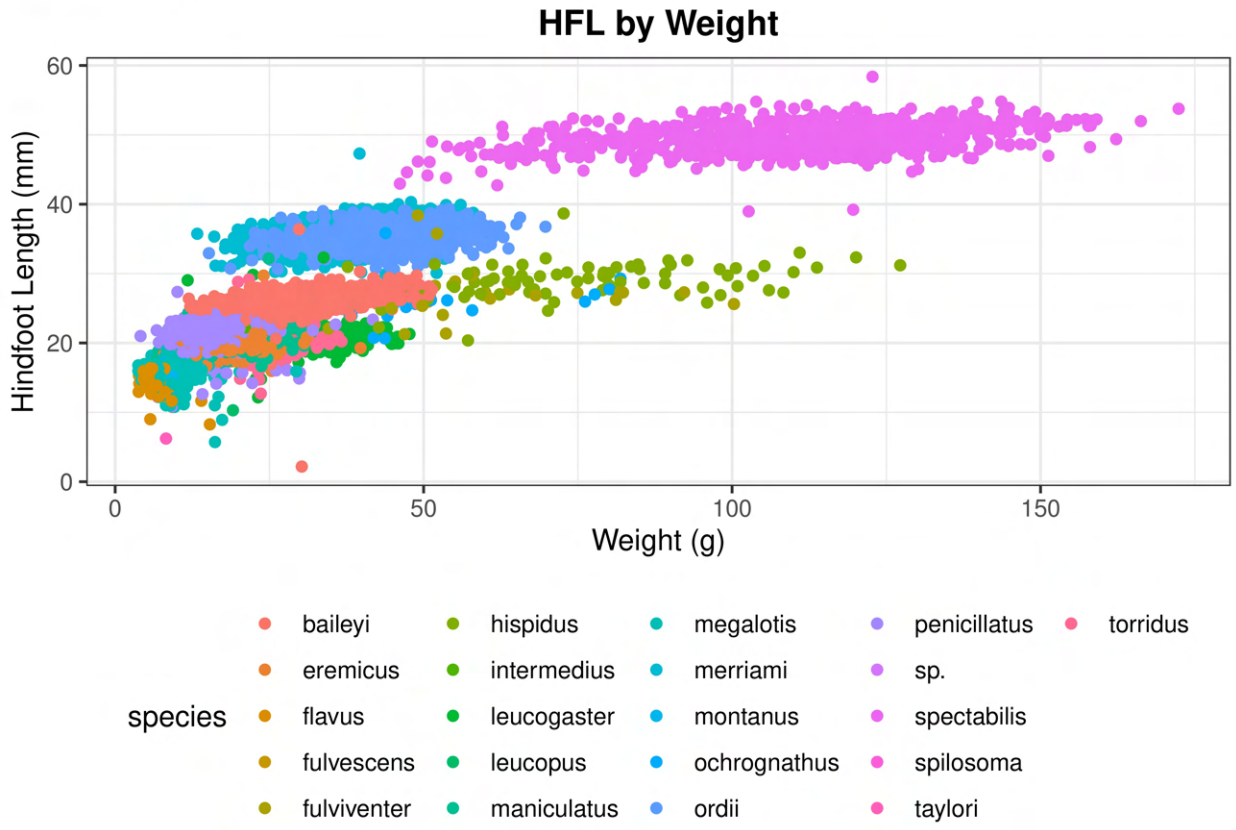
H0 = No statistical significant difference in HFL Averages across different ages, species, sex, and weight

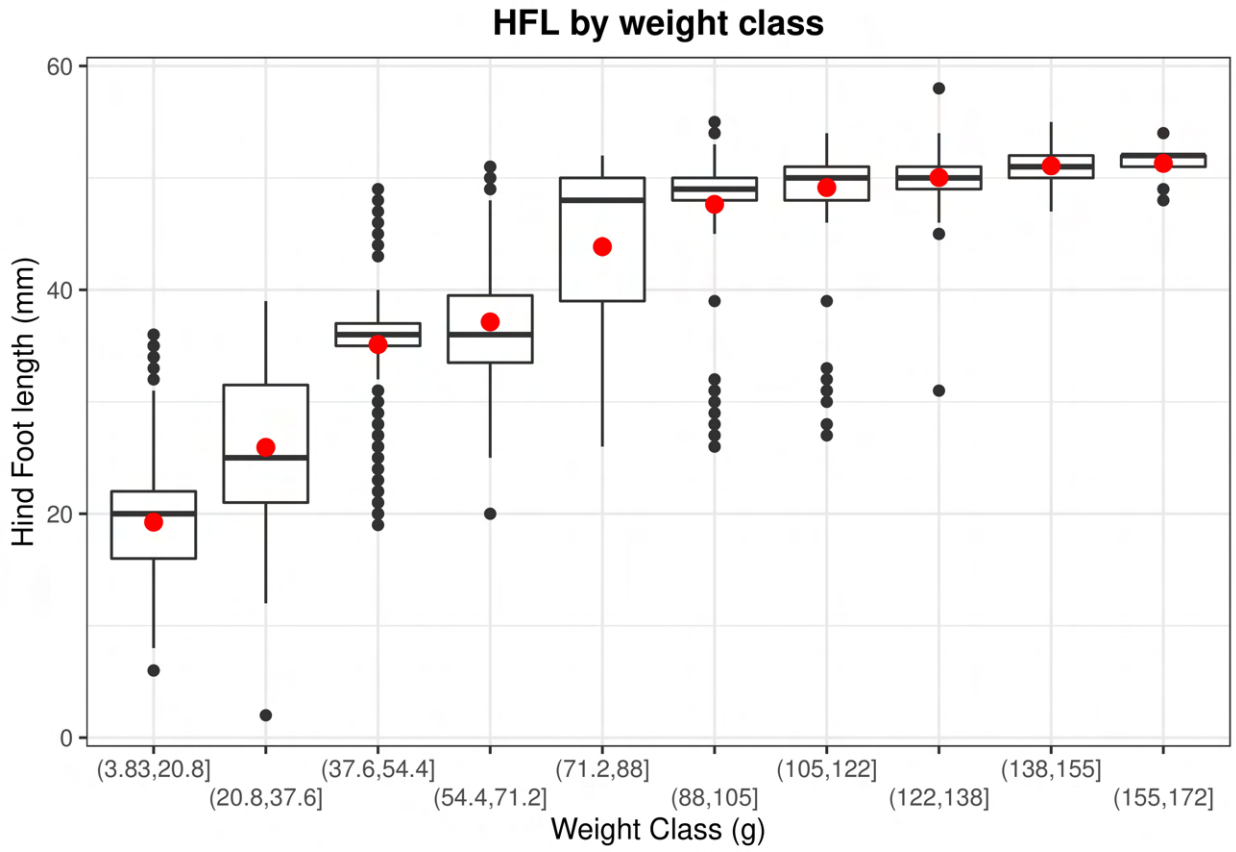
Ha = At least one statistical significant difference in HFL Averages across different ages, species, sex, and weight

Exploratory: Data Visualization

Looking at the initial data visualization, we can see two things. In the first scatter plot, within each species we see somewhat of a positive trend between the weight of the animal and their hind foot length. The second

graph depicts animals in the data set where their weights are subdivided into 10 equally sized binned weight classes. From this we can see that from their mean that there seems to have some sort of relationship between weight and hind foot length.





Exploratory: Modeling

A linear model was created using the following equation:

$$\begin{aligned}HFL =&\beta_0 + \beta_1(AH?) + \beta_2(AS?) + \beta_3(BA?) + \beta_4(CB?) + \beta_5(CM?)+\\ &\beta_6(CQ?) + \beta_7(CS?) + \beta_8(CT?) + \beta_9(CU?) + \beta_{10}(CV?)+\\ &\beta_{11}(DM?) + \beta_{12}(DO?) + \beta_{13}(DS?) + \beta_{14}(DX?) + \beta_{15}(EO?) + \beta_{16}(GS?) + \beta_{17}(NL?) + \beta_{18}(NX?)+\\ &\beta_{19}(OL?) + \beta_{20}(OT?) + \beta_{21}(OX?) + \beta_{22}(PB?) + \beta_{23}(PC?) + \beta_{24}(PE?) + \beta_{25}(PF?) + \beta_{26}(PG?)+\\ &\beta_{27}(PH?) + \beta_{28}(PI?) + \beta_{29}(PL?) + \beta_{30}(PM?) + \beta_{31}(PP?) + \beta_{32}(PU?) + \beta_{33}(PX?) + \beta_{34}(RF?)+\\ &\beta_{35}(RM?) + \beta_{36}(RO?) + \beta_{37}(RX?) + \beta_{38}(SA?) + \beta_{39}(SB?) + \beta_{40}(SC?) + \beta_{41}(SF?) + \beta_{42}(SH?)+\\ &\beta_{43}(SO?) + \beta_{44}(SS?) + \beta_{45}(ST?) + \beta_{46}(SU?) + \beta_{47}(SX?) + \beta_{48}(UL?) + \beta_{49}(UP?) + \beta_{50}(UR?)+\\ &\beta_{51}(US?) + \beta_{52}(ZL?) + \beta_{53}(ZM?) + \beta_{54}(Weight)+\\ &\beta_{55}(Male?) + \beta_{56}(Juvenile?)\end{aligned}$$

From the term estimates, it seems that all of the terms are significant. Moreover, from the model diagnostics looking at the R^2 value of .9813879. This means that the model is really good at capturing the variation within the data. That said, examining the number of terms, it is very clear that there is a very large degree of freedom. We will then evaluate this model to a couple of future models to see if they are better. Additionally, a Type II anova test was performed on the model it was found that Age captured NONE of the variance seen in the data set and was also not significant at the 5% level. Examining the other terms, we see that Sex and Wgt although captures less of the data when compared to species, has a significant p value. The term estimates are placed in the appendix under "Term Estimates for Initial Exploratory Model".

Table 1: Model Diagnostic Data

r.squared	p.value	AIC	BIC	df
0.9813879	0	46521.38	46717.3	24

Table 2: Anova Type II

term	sumsq	df	statistic	p.value
sex	1.447152e+02	1	85.9450281	0.0000000
wgt	2.016761e+03	1	1197.7360850	0.0000000
age	4.699765e-01	1	0.2791148	0.5972905
species_id	3.134999e+05	21	8865.9407802	0.0000000
Residuals	2.326521e+04	13817	NA	NA

Exploratory: Model Comparison

Since the degree of freedom is so high, we want to reduce it to improve our model diagnostics. Particulary this is true with AIC and BIC scores. The first thing we will do with our model is to drop the age variable from the study. Since it is capturing 0% of the data and had no significance. Dropping this variable will have no significant impact on the term estimates. We also want to see if we can improve the amount of data captured by our model. Since the type of species can have interacted effect with weight, we will add this term in. Recall that adding more variables makes the overall model diagnostic worse (more degrees of freedom). So we want to see if the benefits would outweight the cost of degrees of freedom.

Shown below are the model diagnostics for the two new models. Comparing the ageless model with our original, we find that the new model had no changes to the R^2 value and the AIC/BIC scores AIC improved by ~2 pts and BIC improved by ~10 pts. Since the rest of the data (terms and r^2) stayed the same. This model is clearly better. Looking at the model with interaction model and the original, we find that the AIC and BIC score worsened by ~1100 points and the variance captured by the data decreased. From the model comparison, we will move forward with the Age removed from our original model.

Table 3: Model Diagnostic Data for model without Age

r.squared	p.value	AIC	BIC	df
0.9813875	0	46519.66	46708.04	23

Table 4: Anova Type II: Model without Age

term	sumsq	df	statistic	p.value
sex	1.447152e+02	1	85.9450281	0.0000000
wgt	2.016761e+03	1	1197.7360850	0.0000000
age	4.699765e-01	1	0.2791148	0.5972905
species_id	3.134999e+05	21	8865.9407802	0.0000000
Residuals	2.326521e+04	13817	NA	NA

Table 5: Model Diagnostic Data for Interactions

r.squared	p.value	AIC	BIC	df
0.9568822	0	58148.36	58336.75	23

Table 6: Anova Type II: Model with Interactions

term	sumsq	df	statistic	p.value
sex	1.447152e+02	1	85.9450281	0.0000000
wgt	2.016761e+03	1	1197.7360850	0.0000000
age	4.699765e-01	1	0.2791148	0.5972905
species_id	3.134999e+05	21	8865.9407802	0.0000000
Residuals	2.326521e+04	13817	NA	NA

Exploratory: Model Outliers

From the outlier analysis, it looks like there is a large amount of data that are over the threshold for both cooksd and .hat. We decided NOT to remove the outliers. This is because removing the outliers only gave a .01 improvement to the model. That said, the df dropped from 23 to 12. This was because certain species entirely was thought of as an outlier for the model. Removing those species would defeat the purpose of the exploratory study which looks at the relationship between weight and hind foot length.

Exploratory: Final Model and Visual

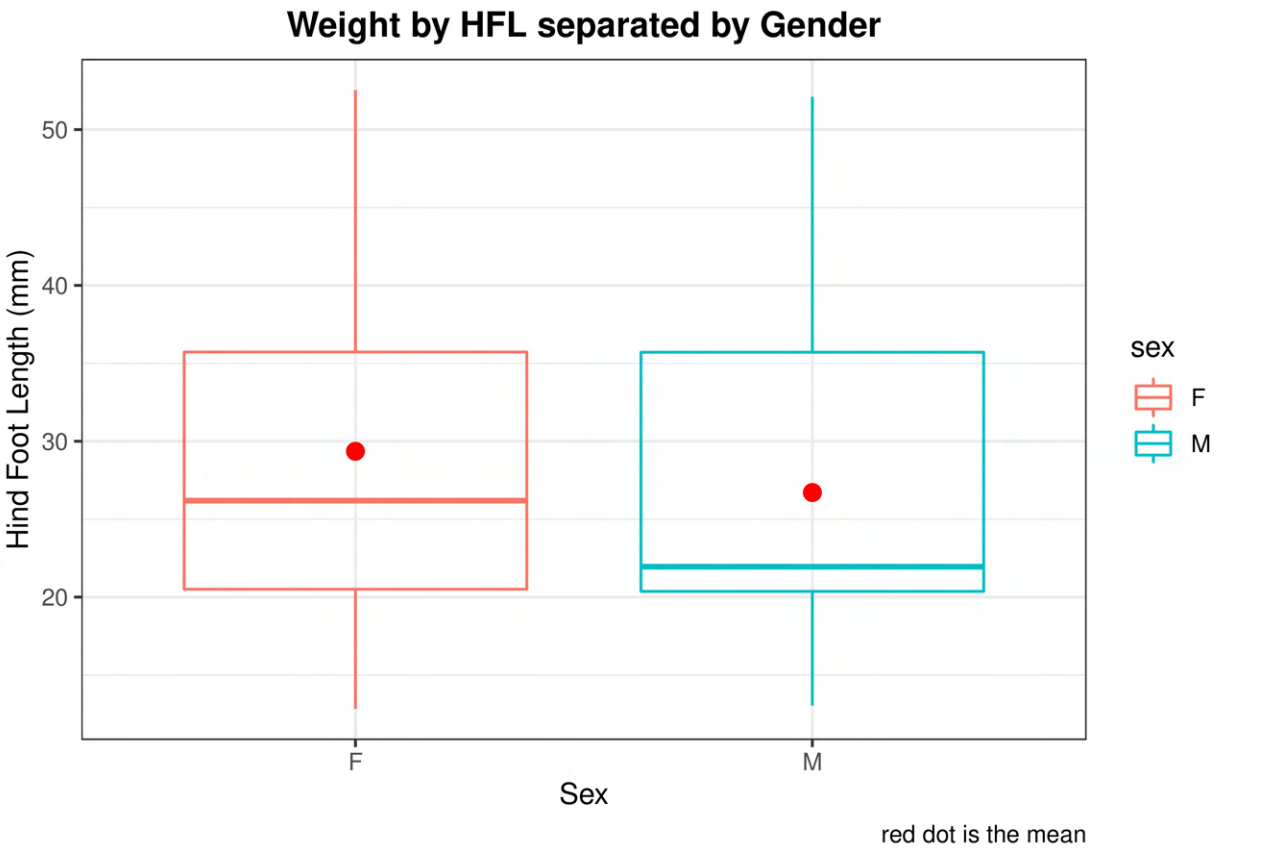
Since we are not removing the models. From our model comparison, the model where age was dropped was the best model. We wil be drawing our conclusions from this model. Shown below are the model data and visual. The terms for the final model is in the appendix.

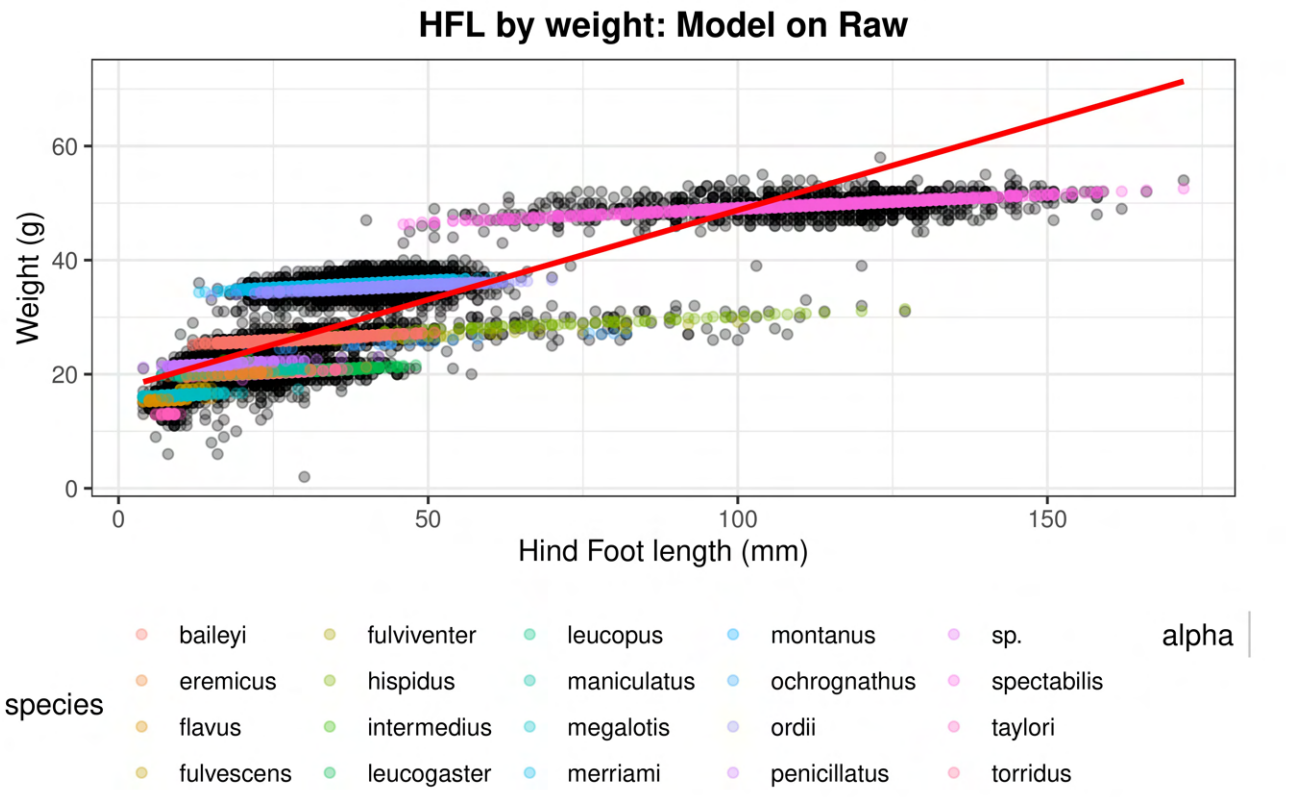
Table 7: R-Squared and Model Diagnostic Scores

r.squared	df	AIC	BIC
0.9813875	23	46519.66	46708.04

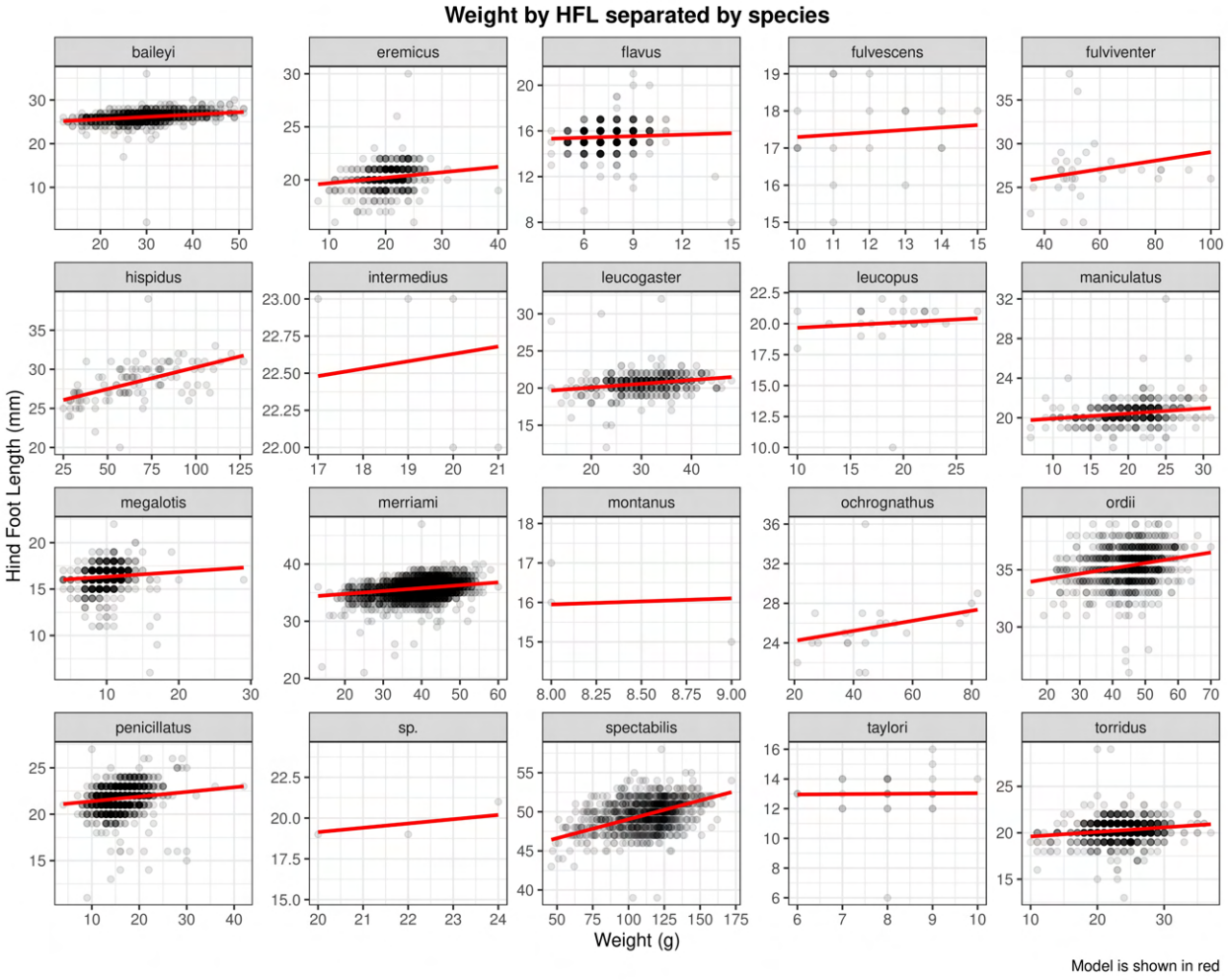
Table 8: Anova Type II Results

	Sum Sq	Df	F value	Pr(>F)
sex	144.7556	1	85.97355	0
wgt	2022.1377	1	1200.99188	0
species_id	313680.4737	21	8871.50980	0
Residuals	23265.6849	13818	NA	NA


Weight by HFL separated by Gender
Hind Foot Length (mm)
50
40
30
20
F
M
Sex
sex
F
M
red dot is the mean



The red line is the general trend across all species



Conclusion

Does weight have an effect on hind foot length?: Examining the graph "HFL by weight" we see that there are clear striations of different species at different weights. They all seem to have a positive trend. When we take into consideration of the weight class to hindfoot length we see a clear positive relationship shown by the red line. To further support this answer, looking at the Anova results taken on the model, we see that weight (wgt) captures a large portion of the variation seen within the graph. Additionally it is significant with F(1,23)[1200.99][P = ~0] It is true however that weight captures less than that of species, but nonetheless it has a impact on the hind foot length. The heavier the animal, the longer its hind foot. Examining the terms page, we see that with a confidence of p<.05, that weight has an approximate impact of .0497 mm increase in hind foot length per gram of weight increase.

Does weight within species have an effect on hind foot length? Examining the graph labeled "Weight by HFL separated by species", the species have been faceted such that each species had its facet window. In some of the species, there were not enough points to get a confident model, however in the species that do, we see that there is a general trend within each species the size of the animal does have an effect on the hind foot length. From the Anova results we see that species captures the largest number of variation seen within the data with a F(21,23)[8871.5][P = ~0].

Does sex have an effect on hind foot length? Examining the graph labeled "Weight by HFL separated by gender", the animals were divided into male and female. Using this box plot we see that males in general have a shorter hind lengths than females. However from our t-tests we find males in general have a larger hind foot length when just accounting for sex. The reason why the box plot shows that females have a larger

hind foot length is that females generally weigh more than males. This weight is influenced in a separate factor explain above leading to the box plot showing females having longer hind foots. When accounting just for sex, at a F(1,23)[85.95][P = ~0], we find that males were .2 mm longer in hind foot when compared to females.

Does hind length change from juvenile to adulthood? This hypothesis couldn't be answered since we dropped age from our study.

HYPOTHESIS II

DATA CARPENTRY

Data manipulation - transformations, etc. Based off of the a-prori, the only columns needed for the data analysis are 'record_id', 'month', 'sex', 'species', 'decimalLatitude', 'age', and 'pregnant,' but some addition columns were kept for context of the other data points. Only data point with 'sex' entered as 'F' (female) were kept, since the study is on pregnancy, logically only females can become pregnant. Mislabeled 'pregnant' data points were changed to 'N' and all NA values were also changed to 'N'. Erroneous 'pregnant' data points were also changed to 'N.' Finally, 'pregnant' was change to a factor because that is the class variable that will be predicted by the other features. Another data frame was created containing only positive 'pregnant' data points.

A-PRORI HYPOTHESIS

PREFACE

"Survival of the Fittest" has become a phrase used in everyday life, even incorrectly in most cases. The origins of this phrase, comes from Darwin's Theory of Evolution that states that an organism with more favorable traits to survive the environment, will produce more offspring and thus their genetics will supersede the genetics of other the other organisms(National Geographic). Based off of this, a mating season implies that there is an ideal time of the year to have and raise offspring. Using this data set, the relationship between the pregnancy rates of different species at different times of the year and latitude coordinates will be investigated.

$$H_o : \beta_{month:decimalLatitude} = 0$$
$$H_A : \beta_{month:decimalLatitude} \neq 0$$

Scientific Hypothesis: An interaction between time of year and distance from the equator influences whether the animal is pregnant.

DATA VISUALIZATION

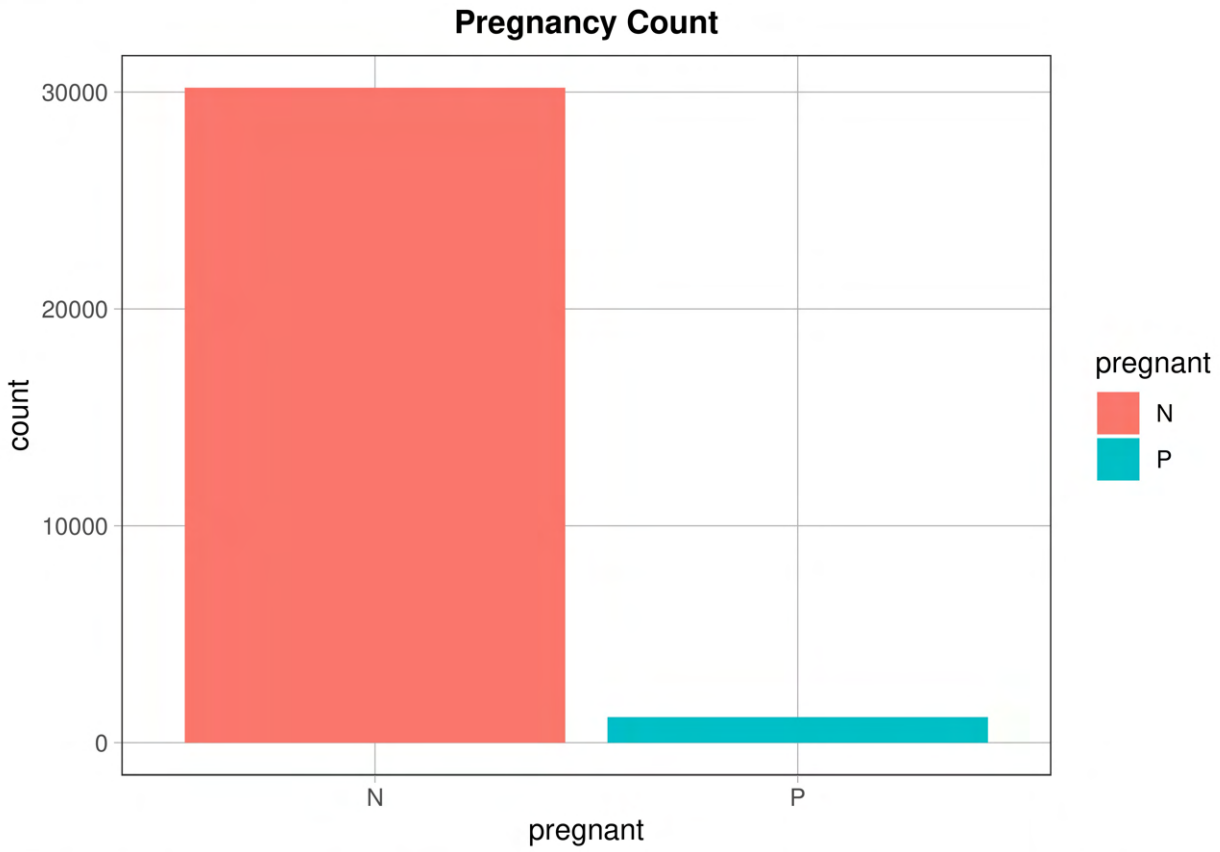


Figure: In this figure, the number of positive and negative pregnancies are displayed. It can be observed that the data is heavily skewed toward negative pregnancy status.

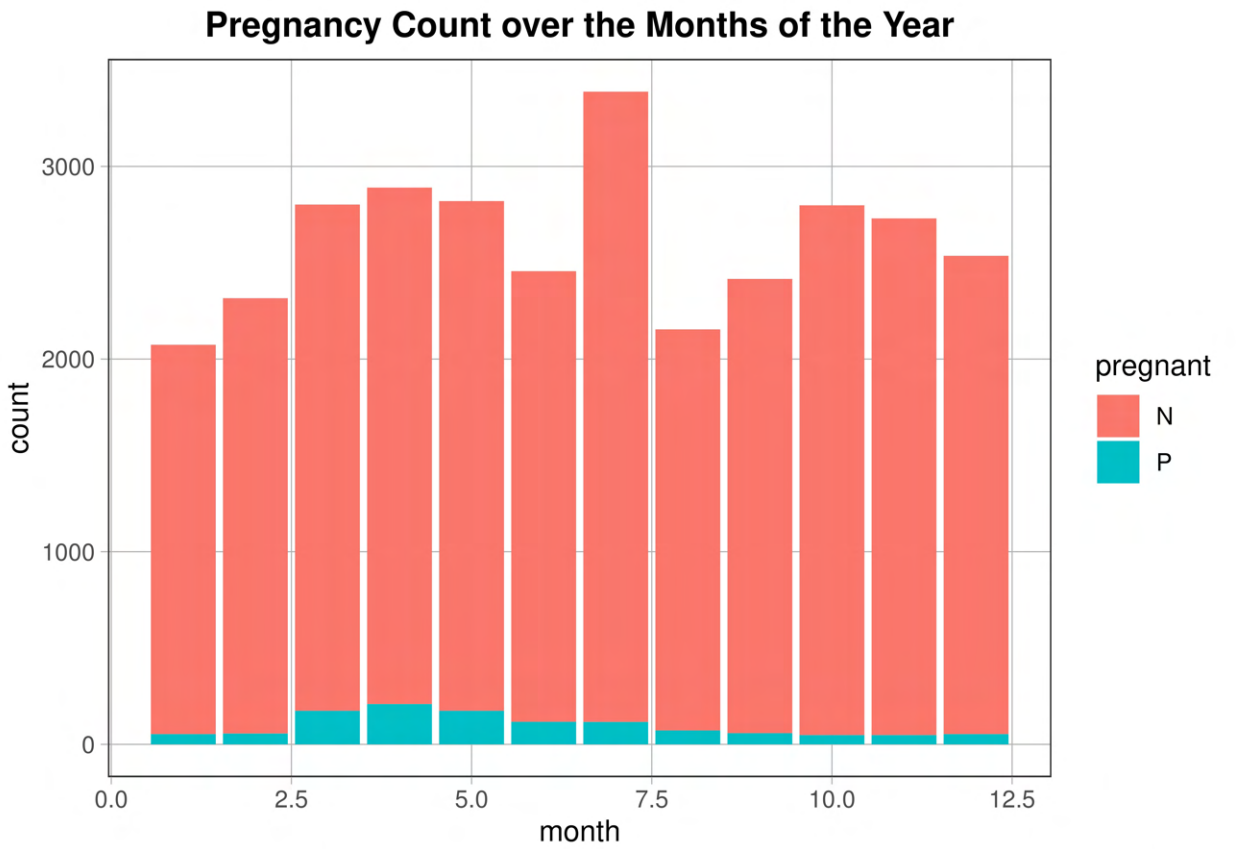


Figure: In this figure, the number of positive and negative pregnancies are displayed according to month. It can be observed that the data is heavily skewed toward negative pregnancy status.There seems to be a slight downward trend of pregnancies starting in the third month of the year.

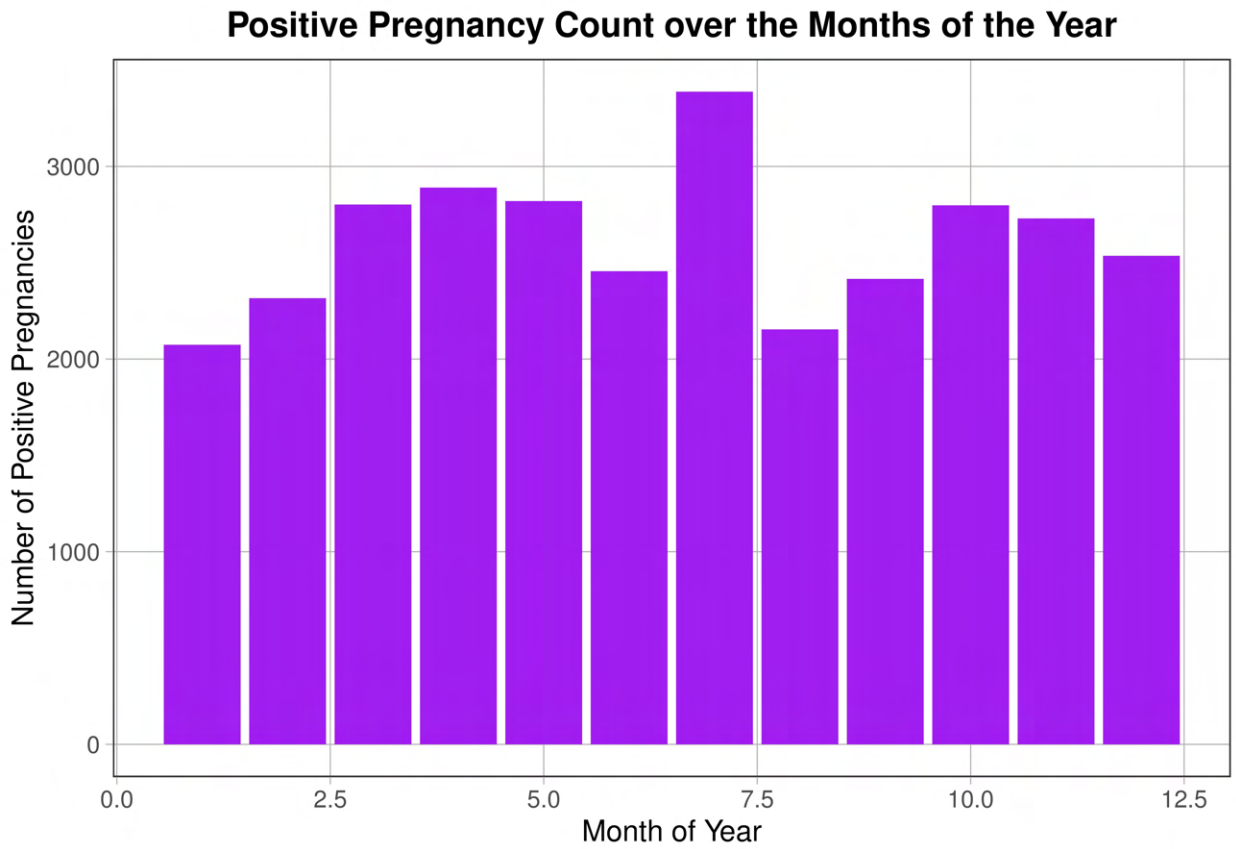


Figure: In this figure, only the number of positive pregnancies are displayed. No real trend can be observed in the data given, but the seven month has the highest amount of pregnancies.


Pregnancy Count over the Months of the Year Seperated by Species
albigula
BA
baileyi
DM
DO
DS
eremicus
flavus
fulvescens
fulviventer
hispidus
leucogaster
leucopus
maniculatus
megalotis
merriami
montanus
OL
ordii
OT
OX
PB
PE
penicillatus
PF
PH
PL
PM
PP
PX
RF
RM
RO
SF
SH
SO
sp.
spectabilis
spilosoma
SS
taylori
torridus
NA
Count
Month of Year
pregnant
N
P

Figure: In this figure, the number of positive and negative pregnancies are displayed according to month and separated by species. It can be observed that the data is heavily skewed toward negative pregnancy status, with only a couple of species displaying any number of positive cases.

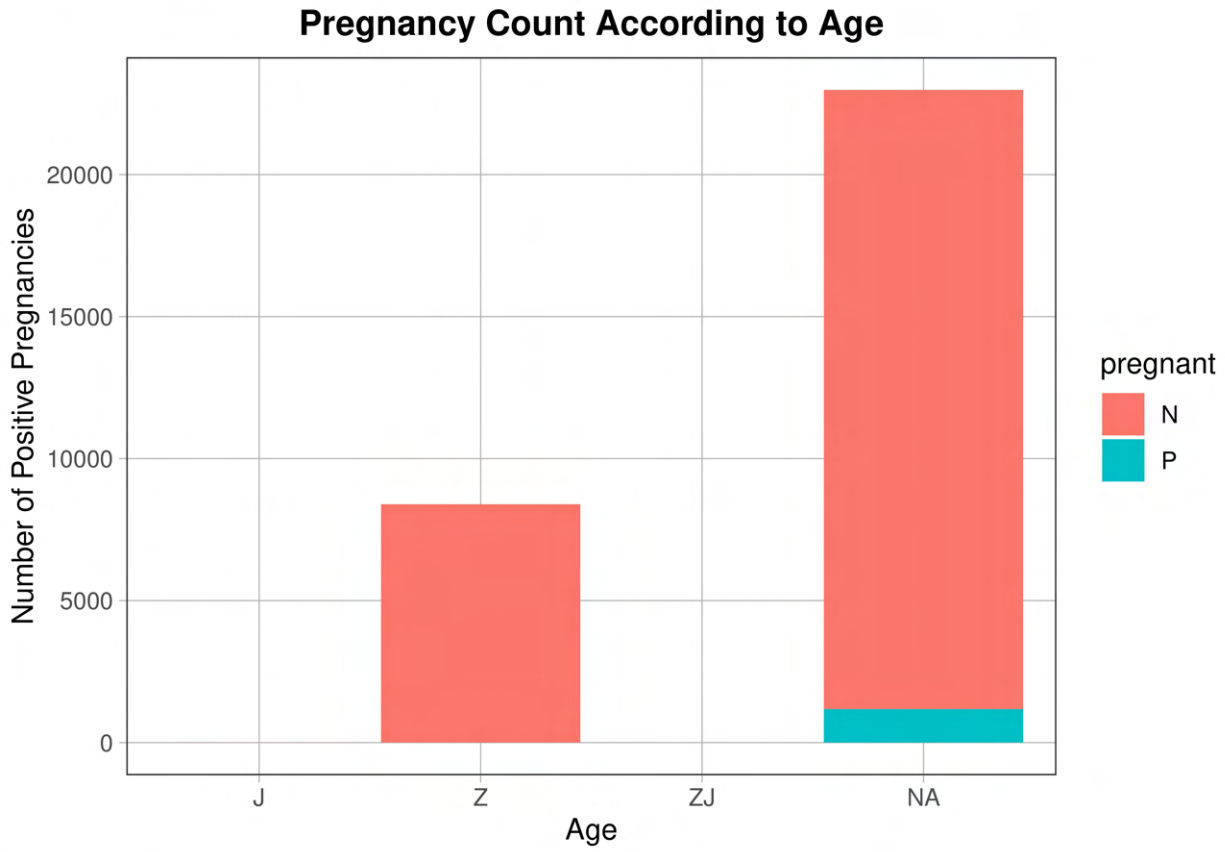


Figure: In this figure, the number of positive and negative pregnancies are displayed according to age. it can be observed that the data is heavily skewed toward negative pregnancy status and an insignificant amount of juvenile animals have pregnancy data.

MODEL DEVELOPMENT

GENERAL EQUATION

$$PregnancyStatus = \beta_0 + \beta_{Species} Species? + \beta_{Age} Age? + \beta_{Month} Month + \beta_{Latitude} Latitdue + \beta_{MonthLatitude} MonthLatitude$$

A PRIORI CONCLUSION

Due to insufficient data, the model could not be run. It should be noted, due to insufficient data, some exploratory hypothesizes could not be investigated and thus are absent from this report. Specifically with 'decimalLatitude' and 'age' data points.

EXPLORATORY HYPOTHESIS

PREFACE

The relationship between time of year and pregnancy status is directly investigated in this exploratory hypothesis. Conditions throughout the time of year fluctuate with the changes in the distance from the sun, so there could be an adaptation to mate during a specific time of year, irregardless of location and species.

$$H_o : \beta_{month} = 0$$
$$H_A : \beta_{month} \neq 0$$

Scientific Hypothesis: Time of year influences whether the animal is pregnant.

DATA VISUALIZATION

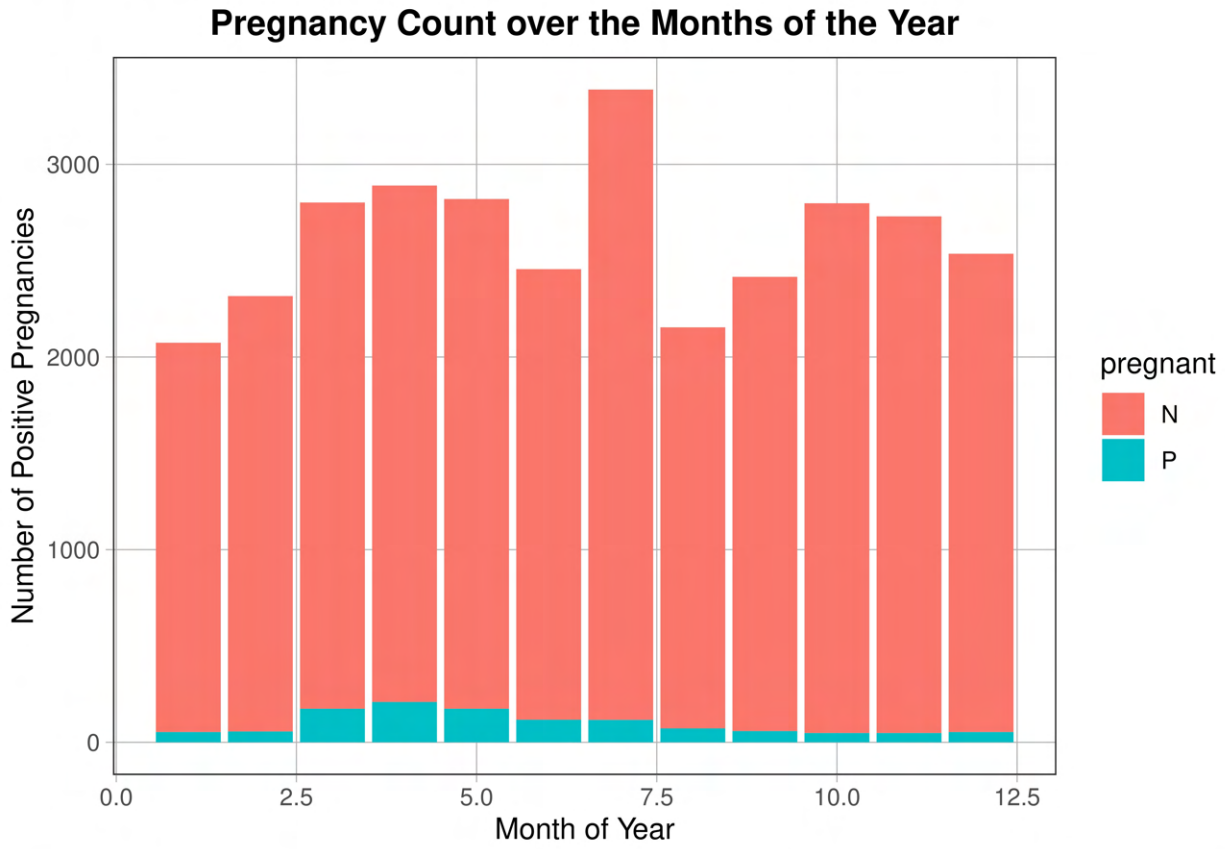


Figure: In this figure, the number of positive and negative pregnancies are displayed according to month. It can be observed that the data is heavily skewed toward negative pregnancy status.There seems to be a slight downward trend of pregnancies starting in the third month of the year.

MODEL DEVELOPMENT

Table 9: Month Hypothesis Coefficients

term	estimate
(Intercept)	-2.7061777
month	-0.0877652

GENERAL EQUATION

$$PregnancyStatus = \beta_0 + \beta_{Month} Month$$

MODEL VISUALIZATION

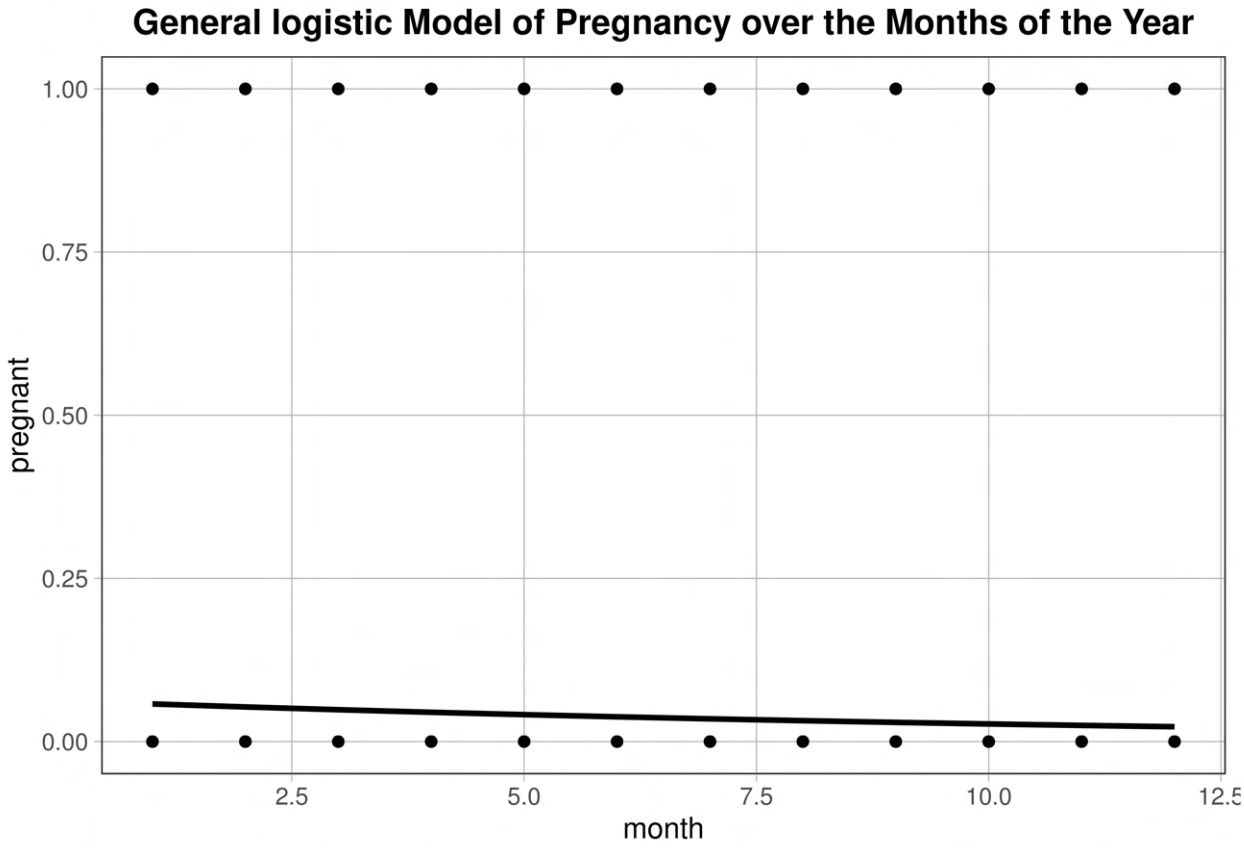


Figure: In this figure, positive pregnancy is displayed as the value 1 and negative pregnancy is displayed as the value zero. The line is the model of the general logistic regression, but it can be observed that the negative pregnancy status pulls the line to a negative results.

ANALYSIS OF MODEL

Table 10: In this table, the predictions of the model using variable cutoff can be observed. All predictions result in 'No,' so a confusion matrix can be not used to evaluate the model because this model can not predict pregnancy status based solely off month.

pred_10	pred_50	pred_90
No	No	No
No	No	No
No	No	No
No	No	No
No	No	No
No	No	No

EXPLORATORY CONCLUSION

Due to a skew towards 'N' in the 'pregnant' class, the model can not be properly evaluated because the model only predicts 'No'.

EXPLORATORY HYPOTHESIS

PREFACE

The relationship between time of year, species, and pregnancy status is directly investigated in this exploratory hypothesis. Conditions throughout the time of year fluctuate with the changes in the distance from the sun, so there could be an adaptation of certain species to mate during a specific time of year, irregardless of location.

$$H_o : \beta_{month} + \beta_{species} = 0$$
$$H_A : \beta_{month} + \beta_{species} \neq 0$$

Scientific Hypothesis: Time of year and species influences whether the animal is pregnant.

DATA VISUALIZATION

Figure: In this figure, the number of positive and negative pregnancies are displayed according to month and separated by species. It can be observed that the data is heavily skewed toward negative pregnancy status, with only a couple of species displaying any number of positive cases.

MODEL DEVELOPMENT

Table 11: Month and Species Hypothesis Coefficients

term	estimate
(Intercept)	-20.0316005
month	-0.0813045
speciesBA	19.2934086
speciesbaileyi	0.0376063
speciesDM	17.4724993
speciesDO	17.5705826
speciesDS	17.1217574
specieseremicus	-0.0447492
speciesflavus	0.0008997
speciesfulvescens	-0.1336597
speciesfulviventer	-0.0430769
specieshispidus	-0.0839488
speciesleucogaster	-0.0548637
speciesleucopus	-0.0872459
speciesmaniculatus	-0.0574294
speciesmegalotis	-0.1035633
speciesmerriami	-0.0426680
speciesmontanus	0.0921267
speciesOL	18.4136085
speciesordii	-0.0704146
speciesOT	18.3716550
speciesOX	0.1560015
speciesPB	17.6741395
speciesPE	19.3686404
speciespenicillatus	0.0250264
speciesPF	18.2808177
speciesPH	17.5162498
speciesPL	18.5328219
speciesPM	19.1530251
speciesPP	17.8532417
speciesPX	-0.3718591
speciesRF	18.8879919
speciesRM	18.8870032
speciesRO	0.0921267
speciesSF	17.7416527
speciesSH	17.2479872
speciesSO	-0.0135543
speciessp.	0.0321564
speciesspectabilis	-0.0281402
speciesspilosoma	0.2785769
speciesSS	0.2785769
speciestaylori	-0.0493209
speciestorridus	0.0235153

GENERAL EQUATION

$$PregnancyStatus = \beta_0 + \beta_{Month} Month + \beta_{Species} Species?$$

MODEL VISUALIZATION

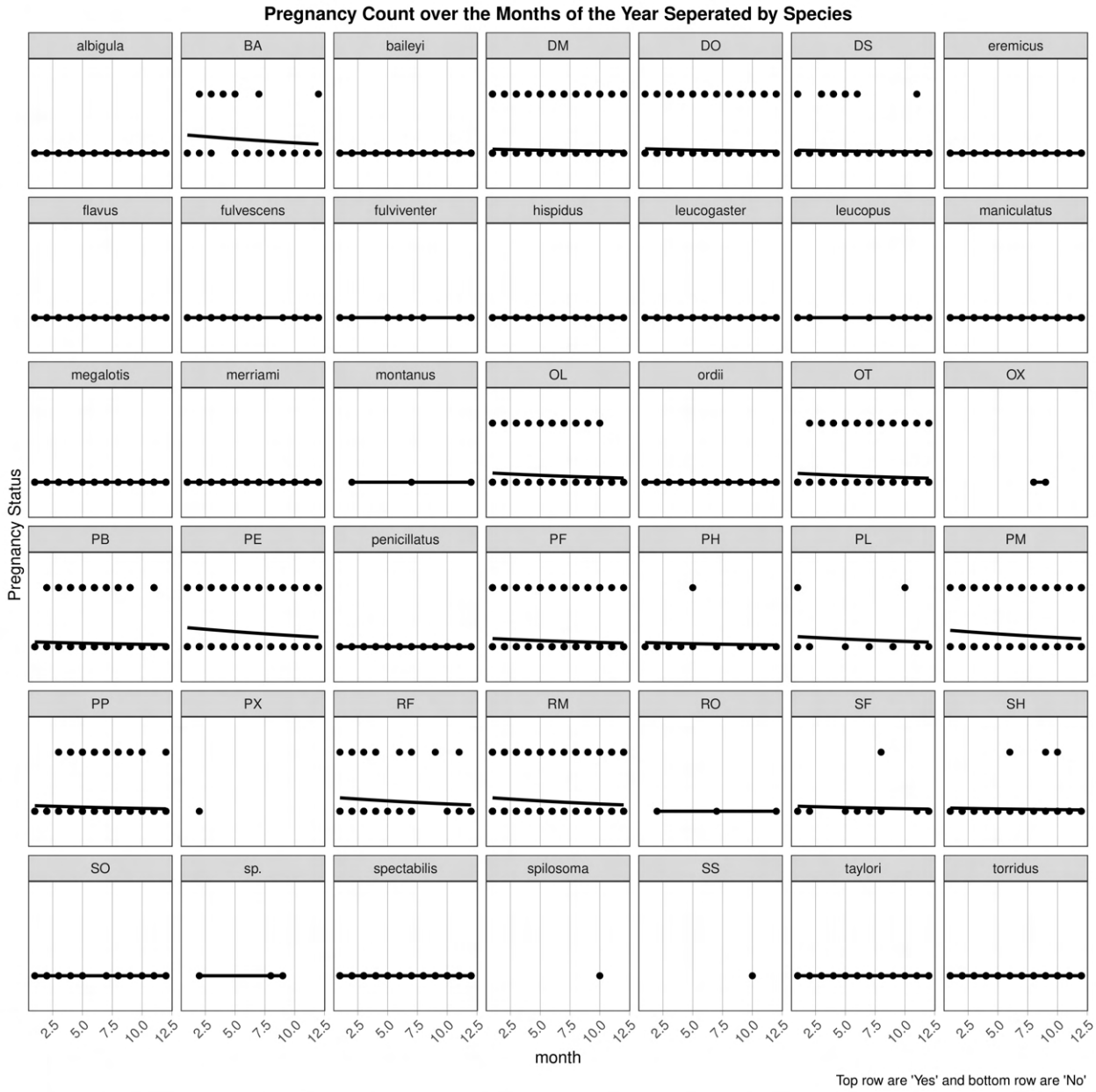


Figure: In this figure, the number of positive and negative pregnancies are displayed according to month and separated by species. It can be observed that the data is heavily skewed toward negative pregnancy status, with only a couple of species displaying any number of positive cases. This results in only some species displaying the line of the general logistic regression model. Within these minimal results, it can be observed that the negative pregnancy status pulls the line to a negative results.

ANALYSIS OF MODEL

Table 12: Model Predictions

pred_10	pred_50	pred_90
No	No	No
No	No	No
No	No	No
No	No	No
No	No	No
No	No	No

Table 13: Evalutation Results from Confusion Matrix

Accuracy	Sensitivity	Specificity
0.8931119	0.4901288	0.9090047

Table: In this table, the predictions of the model using variable cutoff can be observed. All predictions result in 'No,' expect with a value of .1 cutoff, so only this confusion matrix can be used to evaluate the model. Overall, the general logistic regression is skewed toward 'No.' Table: In this table, evaluation results are displayed based off of the confusion matrix produced from the prediction model with a .1 cutoff. Based off of the evaluation, the accuracy and specificity are high, but sensitivity is low.

EXPLORATORY CONCLUSION

In this exploratory hypothesis, the linear regression model produces results with a prediction cutoff of .1. Based off these predictions, a confusion matrix was produced and accuracy, sensitivity, and specificity was evaluated. Though the accuracy and specificity is high the sensitivity is low and this may be due to the fact that the data is heavily skewed towards 'No,' so while the model was able to predict two classes, 'No' was favored and 'No' was a majority of the actual points.

HYPOTHESIS III

DATA CARPENTRY

we created a new data frame to include only the pregnant, age, sex, testes, and vagina variables since those are the variables in question for my hypothesis. we converted all of these above variables into factors since none of them included numerical data. For the testes variable, we removed all values that were not S, M, or R since those were the only values that had meaning. For the vagina variable, we removed all values that were not S, P, or B since those were the only values that had meaning. For the pregnant variable, we removed all values that were not P since those were the only values that had meaning. we also changed all of the single values into their worded meanings to better understand the data. Selected weight, country, and hind foot length for exploratory analysis.

A-PRORI HYPOTHESIS

PREFACE

Given that we are provided with data regarding the status of the testes of male animals and the status of the vagina of the female animals, we believe that pregnancy could possibly be predicted by the status of these sex organs. The age of the female animal could also have an effect on whether the animal is pregnant in the assumption that a juvenile animal would most likely not be pregnant. We also want to explore the time of year on pregnancy since most species have a specific mating season.

Statistical Hypothesis:

$$H_o : \beta_{Vagina:Testes} = 0$$
$$H_A : \beta_{Vagina:Testes} \neq 0$$

Scientific Hypothesis: The interaction between state of testes and state of vagina will affect the number of pregnancies, while controlling for age and month the of year.

DATA VISUALIZATION

Based on visualization of the data and of the graphs, we have found that the testes variable does not have any actual values that align with pregnancy status. This is due to the fact that the data is organized by each individual animal, and since males cannot become pregnant, it makes sense that there are no values that align with a positive pregnancy. Along with that, the age variable does not have any values that align with pregnancy either. We were curious to see if age has an effect on a positive pregnancy value, however, we cannot explore that since there is no data in that category to compare to pregnancy. Given these facts, we cannot conduct the model we originally planned since the model will not run with missing values in these two variables. The model is written below, but will not be run due to these issues.

MODEL DEVELOPMENT

GENERAL EQUATION

$$y_{Pregnancy} = \beta_0 + \beta_{Testes}X_{Testes}? + \beta_{Vagina}X_{Vagina}? + \beta_{Age}X_{Age}?+$$
$$\beta_{Month}X_{Month}? + \beta_{Testes.Vagina}(X_{Testes} * X_{Vagina})?$$

A PRIORI CONCLUSION

Based on the analysis of the above analysis and figures, we can conclude that the A Priori hypothesis does not work and cannot provide the conclusion to our research question that we wanted. The testes variable was not a smart choice to use in this hypothesis since males are unable to get pregnant and since we did not know how the data would look. Secondly, the age variable provided no valuable data points for this hypothesis. While these variables are the reason for the failure of the hypothesis, the vagina and month variables both have valuable data points for us to explore regarding their effects on pregnancy. The hypothesis proposed in the A Priori involved only factor variables where glm using the binomial family is required, so we have decided to steer away from that direction to look at a different hypothesis using the normal general linear model.

EXPLORATORY HYPOTHESIS

PREFACE

After receiving the data set and the failure of the A Priori hypothesis, we have decided to move into another exploratory hypothesis. We want to explore the main effects of hind foot length, age, and month of the year on the weight of the animal along with interaction of month of the year and hind foot length.

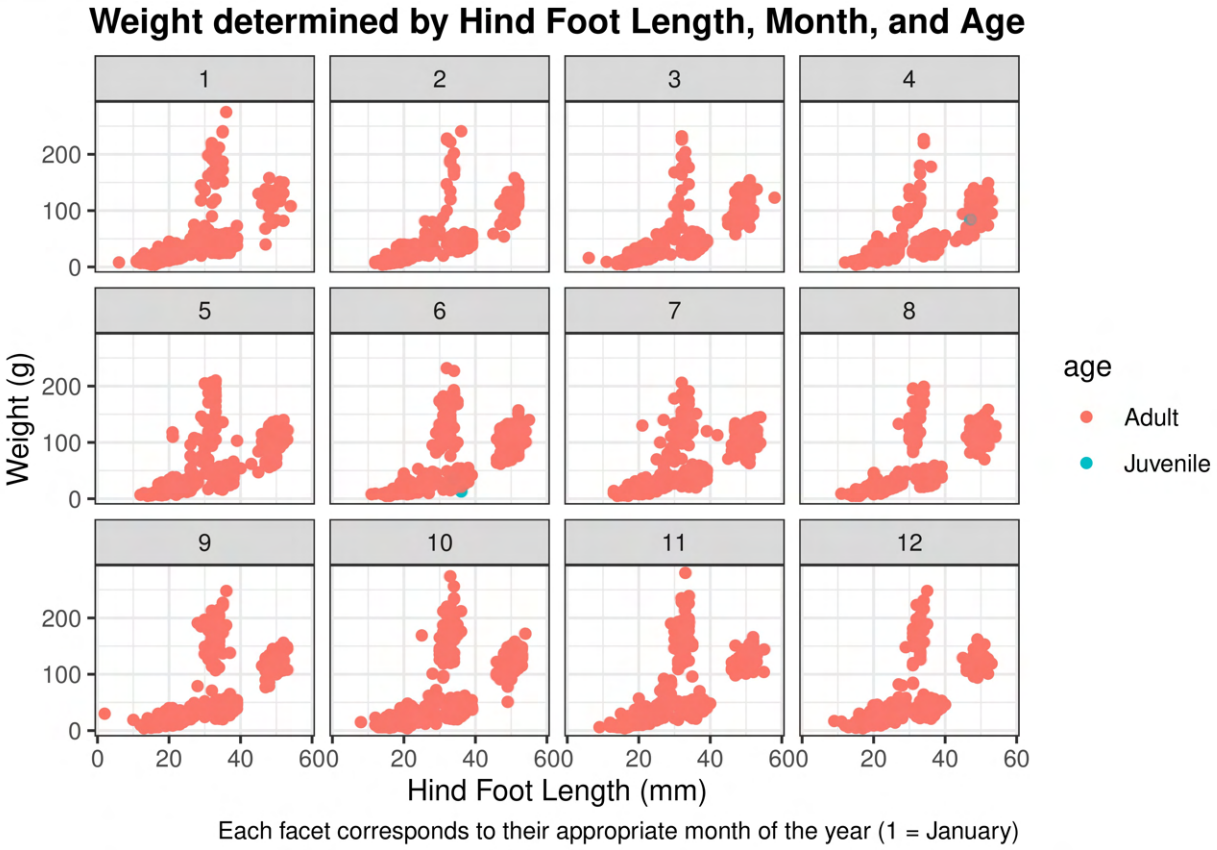
Statistical Hypothesis:

$$H_o : \beta_{Month:HFL} = 0$$
$$H_A : \beta_{Month:HFL} \neq 0$$

Scientific Hypothesis: we expect the interaction of hind foot length and month of the year to have an impact on the weight of the animal holding age constant.

we believe that as hind foot length and age increase, the weight of the animal will increase as well. I also believe that month will have an effect on the animals weight as there might be more food accessibility during warmer months.

DATA VISUALIZATION



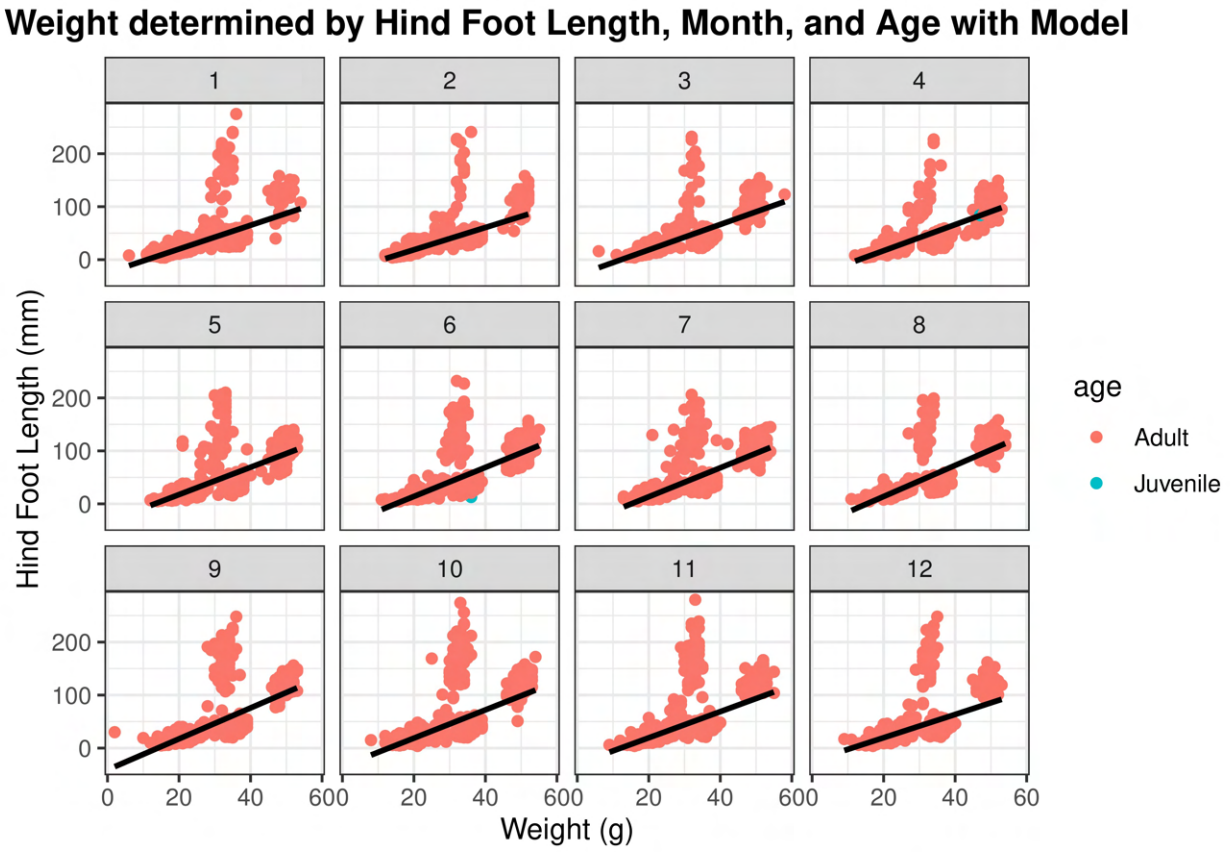
The plot above shows hind foot length on the weight of each animal for each month of the year. Age is also included as a predictor, but as you can see there are little to no juvenile age points. The warmer months appear to have higher amounts of animals with larger weights as predicted, however, each month looks relatively similar. It also appears that as hind foot length increases to around 35, so does the weight of the animal. After hid foot length exceeds 35, weight decreases to around an average of 125.

MODEL DEVELOPMENT

GENERAL EQUATION

$$y_{weight} = \beta_0 + \beta_{hfl}X_{hfl}? + \beta_{month}X_{month}? + \beta_{age}X_{age}?+ \\ \beta_{age:hfl}(X_{age} * X_{hfl})? + \beta_{age:hfl}(X_{month} * X_{hfl})?$$

MODEL VISUALIZATION



ANALYSIS OF MODEL

Table 14: R Squared Value

r.squared
0.4593387

Table 15: Model Beta Terms

term	estimate
(Intercept)	-24.6848629
month2	1.3891138
month3	-4.7236342
month4	-7.7573985
month5	-9.2973267
month6	-16.1038339
month7	-16.5496565
month8	-20.1318713
month9	-16.1976218
month10	-9.8337058
month11	-5.2819967
month12	0.3959088
hfl	2.2402072

term	estimate
ageJuvenile	-20.6855919
month2:hfl	-0.1395945
month3:hfl	0.1558881
month4:hfl	0.2276761
month5:hfl	0.3381113
month6:hfl	0.5038772
month7:hfl	0.4938006
month8:hfl	0.6981745
month9:hfl	0.6753353
month10:hfl	0.4252787
month11:hfl	0.2296020
month12:hfl	-0.0458159

Table 16: Type 3 Sum of Squares Table

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	81185.055	1	121.4913	0.0000
month	66821.527	11	9.0906	0.0000
hfl	548733.273	1	821.1646	0.0000
age	1278.694	1	1.9135	0.1666
month:hfl	82877.223	11	11.2749	0.0000
Residuals	9601241.642	14368	NA	NA

Analyzing the sum of squares table, we can conclude that there is some degree of statistical significance regarding month and hind foot length on the weight of the animal. Each of the corresponding F statistics are very large with p-values well below the alpha value of 0.05, showing statistical significance. Type III: F(11, 14368) = [11.2749], p = [0.0000] These results showed that there is a statistically significant interaction between the effects of hind foot length and month of the year on the weight of the animal. These terms also capture the most data, and result in a majority of the dataset's variance as seen in the high sum of squares regression values of both terms and the interaction of the terms. Along with this analysis, the r squared value is below 0.5 which can indicate that this model is not a great fit for the data. This is most likely due to the higher weights that are in the middle of the hind foot lengths making the data not fully linear.

OUTLIER CHECK

Based on the graph above, there don't appear to be significant outliers that would affect the model.

HYPOTHESIS IV

DATA CARPENTRY

(NOTE: Data carpentry was done in a manner that creates one data set with the necessary information for completion of both the a priori and exploratory hypotheses.)

Outline for Hypothesis IV Data Carpentry: 1) Reduce the parent data set to just the columns needed for our particular a priori and exploratory analysis - month, day, year, decimalLatitude, taxa, record_id, species, country, locality 2) Formulate a separate data set to keep track of the various localities where the data was pulled from. Manually, these locations were looked up on google maps, with care to be as precise as possible, and the resulting latitude input into this locality reference table. This allows us to have a locality reference set that can be used for increasing the number of decimalLatitude points we have. **Note: The locality,

'valley' was determined to be too vague to give a latitude therefore it was assigned an NA and removed from the data set in the following step of data carpentry. 3) Merge the locality reference set with the original reduced data set for the hypothesis. Yield increased number of decimalLatitude points in the set. 4) Because the dependent variable in these analyses is decimalLatitude, we can omit all rows that have an NA in this column. An NA in the decimalLatitude column can not safely be assumed to be a particular latitude. 5) Delete more extraneous columns that are not needed for analysis. 6) Merge with the species reference data set to get the taxa information we need to determine which entries are birds, a major qualifier for us. 7) Remove extraneous columns and correct data types for appropriate graphing. 8) Normalize the day and month data columns to yield a variable for proportion of the year. Based on the day, month, and leap year status, calculate the proportion of the year that has passed at the time of data collection. This yields a better understanding of the time line of migration and makes graphing a time variable on the x-axis more palatable. This new scale spans from January 1st (0) to December 30th (1.0). 8) Get rid of day and month columns that are now extraneous. 9) Set record ID and species to factors, needed for graphing and developing models properly.

The proportion of the year and decimalLatitude do not appear to be colinear based on an exploratory analysis of the two variables plotted against one another. Graphs of this data can be seen in the supplemental materials section.

A-PRORI HYPOTHESIS

PREFACE

Birds follow migratory patterns often moving south for the winter months to have offspring and protect themselves from freezing temperatures. The ecological study of this practice has occurred for many centuries. Our group plans on using the given ecological data to see if we can detect these seasonal patterns of movement among different species. We have chosen to divide the bird taxa into species because various species favor various latitudes and thus temperatures so averaging them across all birds might yield unfavorable results.

Statistical Hypothesis:

$$H_o : \beta_{Latitude:Species} = 0$$
$$H_A : \beta_{Latitude:Species} \neq 0$$

***If the beta above is 0, that element has no impact on the model. If it is anything else, it does have an impact on the final model.

Scientific Hypothesis: It is commonly understood that birds follow migratory patterns with the changing seasons that often cause them to move latitudinally. With the data provided, we expect to be able to predict the latitudinal coordinate of birds dependent on an interaction between time of the year and species.

MODEL DEVELOPMENT & DATA VISUALIZATION

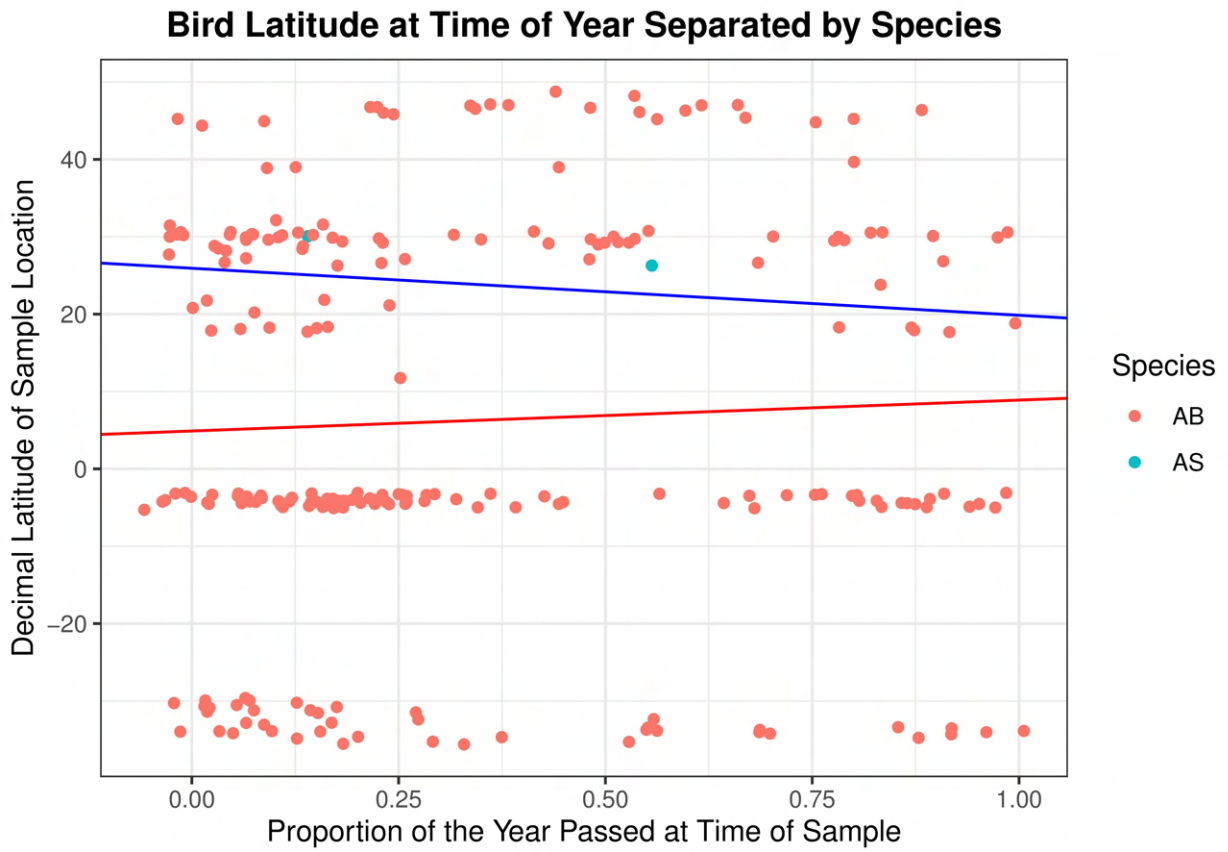
The model for this hypothesis will explore the contributions of species and time of year on the latitude of a bird. Of important note, our model includes the interaction between species and proportion of the year which we will be examining to see if various species have different migration patterns that occur on different time scales through the year.

GENERAL EQUATION

$$decimalLatitude = \beta_0 + \beta_{AS}(AS)? + \beta_{propYear}(propYear) +$$
$$\beta_{propYear,AS}(propYear)(AS)? + \beta_{propYear,AB}(propYear)(AB)?$$

*** β_0 in this case represents the AB species of bird

MODEL VISUALIZATION



ANALYSIS OF MODEL

GENERAL EQUATION with Calculated Coefficents Coefficients for this linear model were retrieved from the tidy of our linear model outlined in the above section.

$$decimalLatitude = 9.22 + 21.6 * (AS)? + -1.07 * (propYear) + \\ -5.01 * (propYear)(AS)?$$

R-squared evaluates how well our determined model fits the original data. Typically, the higher the r-squared, the better fit the model is for that particular spread of data. For this model, r-squared is 0.00496. This low value means that our model neglects to account for variance in our data. Of note, this is data randomly collected by humans on animals. The nature of this data means that we will have some natural variation because these things are harder to predict than finite mechanical processes. Examining our residuals plot (QQ), we can see that our data has some clear groupings that cause certain volumes of residuals to be higher, contributing to the low r-squared. The QQ plot for these graphs can be found in the supplementary materials section.

We have decided to run a III type Anova on our model for this hypothesis, as we are primarily asking questions about the interaction term for this hypothesis. The type III Anova respects marginality. The ANOVA focuses on comparing the effects of species and time of year on latitude. The ANOVA reveled that there was not a statistically significant difference in latitude between the two groups of species at certain times of the year (F(1,235) = [0.0051], p = 0.9433). The P-value is greater than the determined α value of 0.05, therefore we fail to reject the null hypothesis. Additionally of note, none of the main effects for this model have significance either, with p-values greater than our pre-determined α.

Table 17: Coefficients for the Linear Model of Latitude by Time with Species

term	estimate	std.error
(Intercept)	4.892638	2.420465
species_idAS	25.933511	28.853864
propYear	3.997308	5.327895
species_idAS:propYear	-10.076400	67.760528

Table 18: R-squared for the Linear Model of Latitude by Time with Species

r.squared
0.0089972

Table 19: ANOVA (III) Results for the Linear Model of Latitude by Time with Species

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	2646.2192	1	4.0859073	0.0443760
species_id	523.1810	1	0.8078201	0.3696855
propYear	364.5535	1	0.5628906	0.4538484
species_id:propYear	14.3217	1	0.0221135	0.8819131
Residuals	152196.6733	235	NA	NA

OUTLIER ELIMINATION

*** Outlier graphs can be found in the supplemental section for the hypothesis IV.

I am of the opinion that the outliers that were shown in the Cooks distance and hat plot models are not actually outliers despite their outstanding values on these two graphs. Due to the nature of the data being 'grouped' regionally, with major volumes of points taken in the northern hemisphere, the model we produced obviously points out that the negative latitude points in Australia are 'outliers'. In comparison to the NA points, there is a much smaller volume of Australian points that brings up their qualifiers as outliers in this model development. Due to the way the problem was worded, not taking these rough 'groupings' into account, I don't think we can justifiably remove these points from the larger data set. Instead, I propose that we address these 'groupings' in the exploratory analysis to see if the data can be divided regionally to yield a better understanding of the latitude change in bird migration over time. * NA = North America

EXPLORATORY HYPOTHESIS I

PREFACE

Due to the problems faced in outlier analysis and data visualization in the a priori analysis above, our team has detected that there are some regional groupings in the data that are dependent on the data collectors home locations. We believe that grouping data more tightly by country or even region will yield a better understanding of the original question on if bird latitude (migration patterns) depends on the time of year.

In addition, the data set is going to be reduced further as only two points were supplied for the AS species. This volume is not sufficient enough to draw valid hypothesis. As such, the data set will be reduced further to only examine the AB species of bird.

Statistical Hypothesis:

$$H_o : \beta_{PropYear:Country} = 0$$
$$H_A : \beta_{PropYear:Country} \neq 0$$

***If the beta above is 0, that element has no impact on the model. If it is anything else, it does have an impact on the final model.

Scientific Hypothesis: The interaction between country and time of year has a significant impact on the predictive capability of the linear model for the latitude. As populations of birds are not able to teleport to different localities, restricting their analysis by country could be valuable in determining if their latitude changes with time of the year.

Summary of Exploratory Data Carpentry

1) Reduce the data set to remove the AS species type, leaving only the AB species. This is done because we found that we only have two AS species entries in the set which is not enough to draw conclusions from.
2) Divide up the original data set into regional categories by country. Adjust the identifiers for country with appropriate alpha-2 standard country labels.
3) Change the data type for the country column to a factor.
4) Outlier studies have determined that we can omit the single India point from our data set. With only one point, we cannot determine an appropriate equation to model this data, therefore we reccommend removing it.

H4 Exploratory 1 Model Formation + Visualization

GENERAL EQUATION:

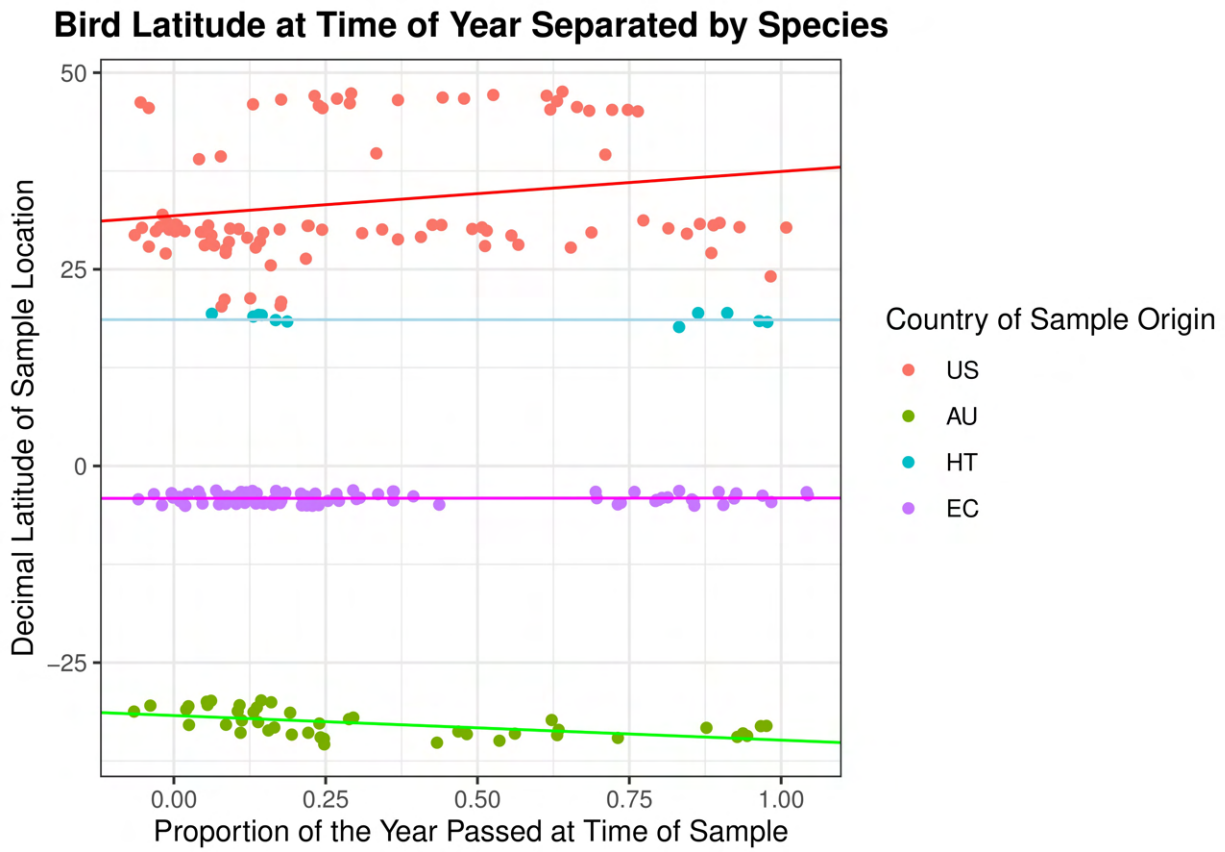
$$\begin{aligned} decimalLatitude = \beta_0 + \beta_{PropYear}(PropYear) + \beta_{AU}(AU)? + \beta_{HT}(HT)? \\ +\beta_{EC}(EC)? + \beta_{AU,PropYear}(PropYear)(AU)? \\ +\beta_{HT,PropYear}(PropYear)(HT)? + \beta_{EC,PropYear}(PropYear)(EC)? \end{aligned}$$

*** β_0 in this case represents the United States

A secondary model without the addition of interaction terms was included for comparison. The equation for that alternative model is as follows:

$$\begin{aligned} decimalLatitude = \beta_0 + \beta_{PropYear}(PropYear) + \\ \beta_{AU}(AU)? + \beta_{HT}(HT)? + \beta_{EC}(EC)? \end{aligned}$$

*** β_0 in this case represents the United States



H4 Exploratory 1 Model Analysis

GENERAL EQUATION WITH FILLED IN COEFFICIENTS:

$$\begin{aligned} decimalLatitude = 31.8 + 5.63 * (PropYear) + -63.5 * (AU)? + -13.2 * (HT)? \\ + - 35.9 * (EC)? + -8.76 * (PropYear)(AU)? \\ + - 5.66 * (PropYear)(HT)? + -5.60 * (PropYear)(EC)? \end{aligned}$$

R-squared gives us a fairly good idea of the amount of variation accounted for by the model. For model 1, which is the model including interactions, the R^^2 was determined to be 0.9629. Model II, which is the model that does not include interactions, the R^^2 was 0.9613. Both of these models yield R^^2 vals that are significantly higher than the a priori model, indicating that they better capture the variation in the data set. Comparing these results of the two models, the model with interactions improves our understanding of the data by having a higher R^^2, but not by a large margin. Further examination of AIC and sum of squares will be able to tell us the value of the interaction term.

Two ANOVA tests were performed to analyze the effects of proportion of the year and country of sample on recorded latitude. One of these ANOVA's was run on the model with interactions while the other was run on the model with interactions. A type II ANOVA was used for the model without interactions while a type III was used for the model with interactions, chosen to respect marginality. Type III (Model I): F(3, 228) = [3.308981], p = [0.0209287] These results showed that there is a statistically significant interaction between the effects of country of recording and time of year on the latitude recorded. This conclusion was reached using an $\alpha = 0.05$, an accepted standard. Outside of the hypothesis, main effects for proportion of the year and country were significant in the Type III model while proportion of the year was not significant in the Type II model. F statistics and p-values can be found in the tables below.

Comparison of the performance of the two models was done with AIC. AIC for model I with interactions

is 1437.399 with 9 degrees of freedom. AIC for model I w/o interactions is 1441.457 with 6 degrees of freedom. From this analysis in addition to the the r-squared values above, it is safe to say that the interaction between country and time of year is not significant in capturing the variation of our data. The model without interactions has a high r-squared in addition to a higher AIC with less degrees of freedom. This tells us that the best model for capturing our data on latitude is country and time of year without interactions.

Table 20: Coefficients for the Linear Model of Latitude by Time with Country

term	estimate	std.error
(Intercept)	31.808477	0.7900095
propYear	5.631357	1.7808146
countryAU	-63.532626	1.3292752
countryHT	-13.201299	2.4255198
countryEC	-35.922970	1.0952593
propYear:countryAU	-8.758629	3.0112346
propYear:countryHT	-5.656287	4.0729416
propYear:countryEC	-5.598470	2.4553521

Table 21: R-squared for the Linear Model of Latitude by Time with Country w/ Interactions

r.squared
0.9629209

Table 22: R-squared for the Linear Model of Latitude by Time with Country w/o Interactions

r.squared
0.9613065

Table 23: ANOVA (III) Results for the Linear Model of Latitude by Time with Country

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	40211.0478	1	1621.142242	0.0000000
propYear	248.0350	1	9.999740	0.0017784
country	62254.5521	3	836.614892	0.0000000
propYear:country	246.2294	3	3.308981	0.0209287
Residuals	5655.3451	228	NA	NA

Table 24: ANOVA (II) Results for the Linear Model of Latitude by Time with Country without Interactions

	Sum Sq	Df	F value	Pr(>F)
propYear	42.95776	1	1.681457	0.1960248

	Sum Sq	Df	F value	Pr(>F)
country	146248.11614	3	1908.152636	0.0000000
Residuals	5901.57450	231	NA	NA

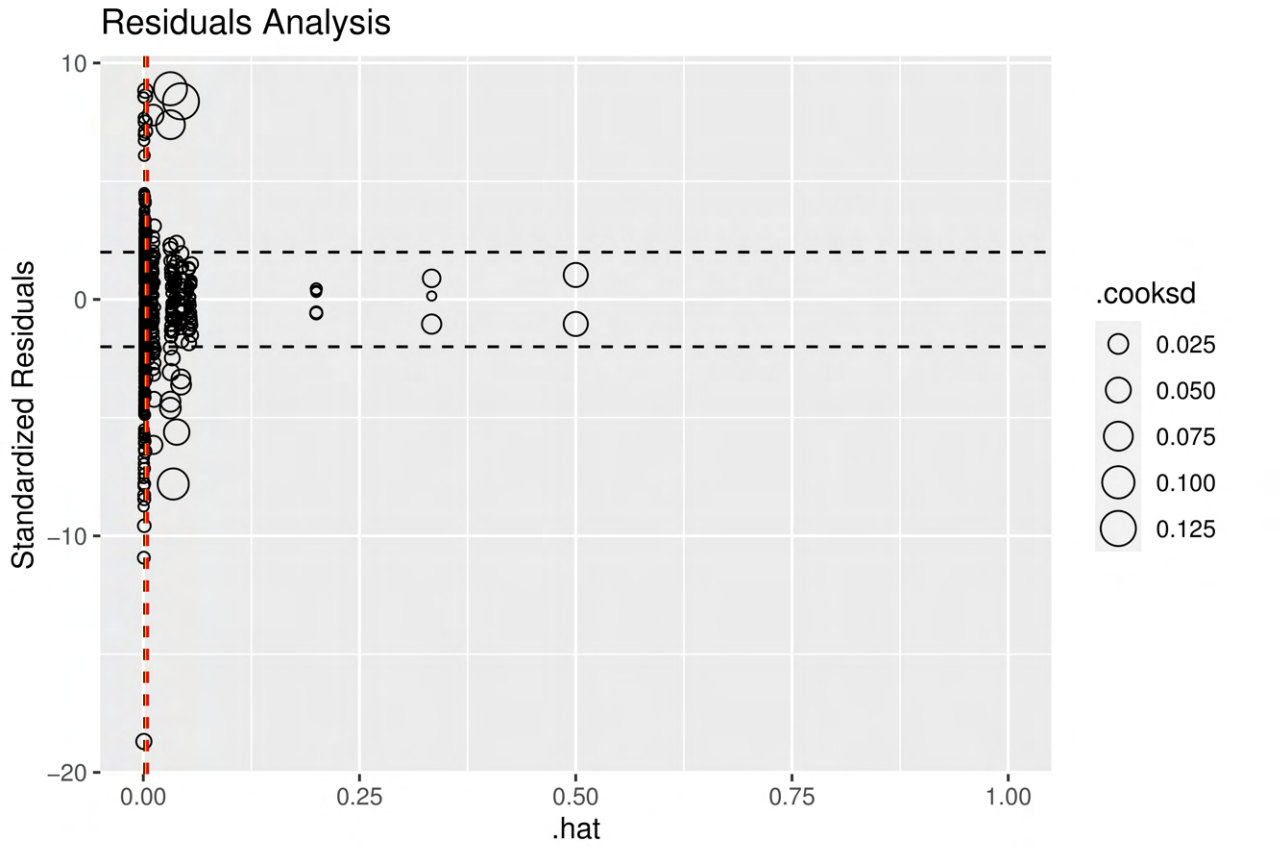
Table 25: AIC Comparison of Latitude Models - One with Interaction and the Other Without

	df	AIC
h4_ex1_model	9	1437.399
h4_ex1_mod2	6	1441.457

Supplemental Data

HYPOTHESIS I

OUTLIERS



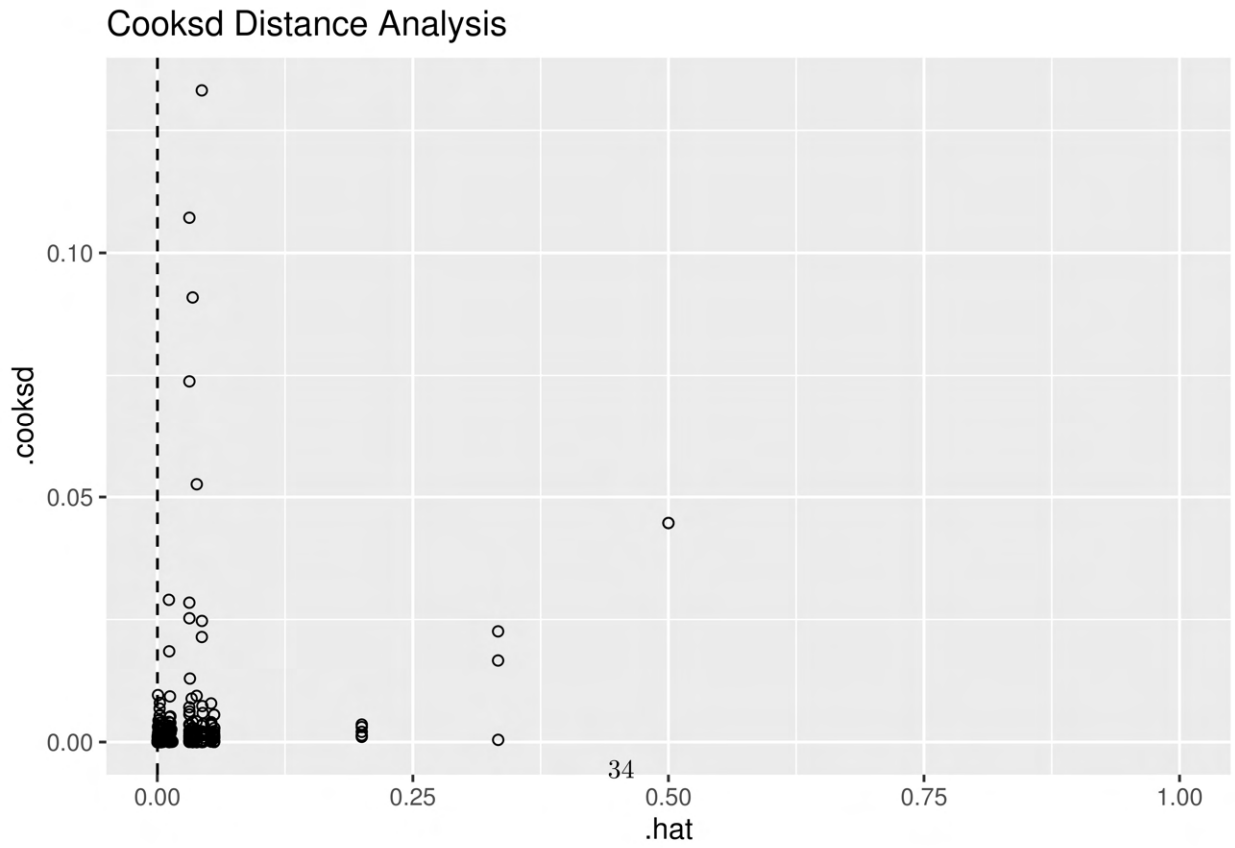


Table 26: Terms for Final Model (ageless)

term	estimate	p.value
(Intercept)	12.5275565	0.0000000
sexM	0.2094474	0.0000000
wgt	0.0497536	0.0000000
species_idDM	21.1851627	0.0000000
species_idDO	20.5080917	0.0000000
species_idDS	31.4561962	0.0000000
species_idOL	6.4435896	0.0000000
species_idOT	6.4509909	0.0000000
species_idOX	6.1186631	0.0000000
species_idPB	12.0175553	0.0000000
species_idPE	6.5351425	0.0000000
species_idPF	2.4978004	0.0000000
species_idPH	11.9313807	0.0000000
species_idPI	8.8977761	0.0000000
species_idPL	6.4767423	0.0000000
species_idPM	6.7831444	0.0000000
species_idPP	8.2679329	0.0000000
species_idPX	5.4773714	0.0000347
species_idRF	4.2549773	0.0000000
species_idRM	3.1892348	0.0000000
species_idRO	2.9181985	0.0002267
species_idSF	11.4449642	0.0000000
species_idSH	12.5702619	0.0000000
species_idSO	10.6486900	0.0000000

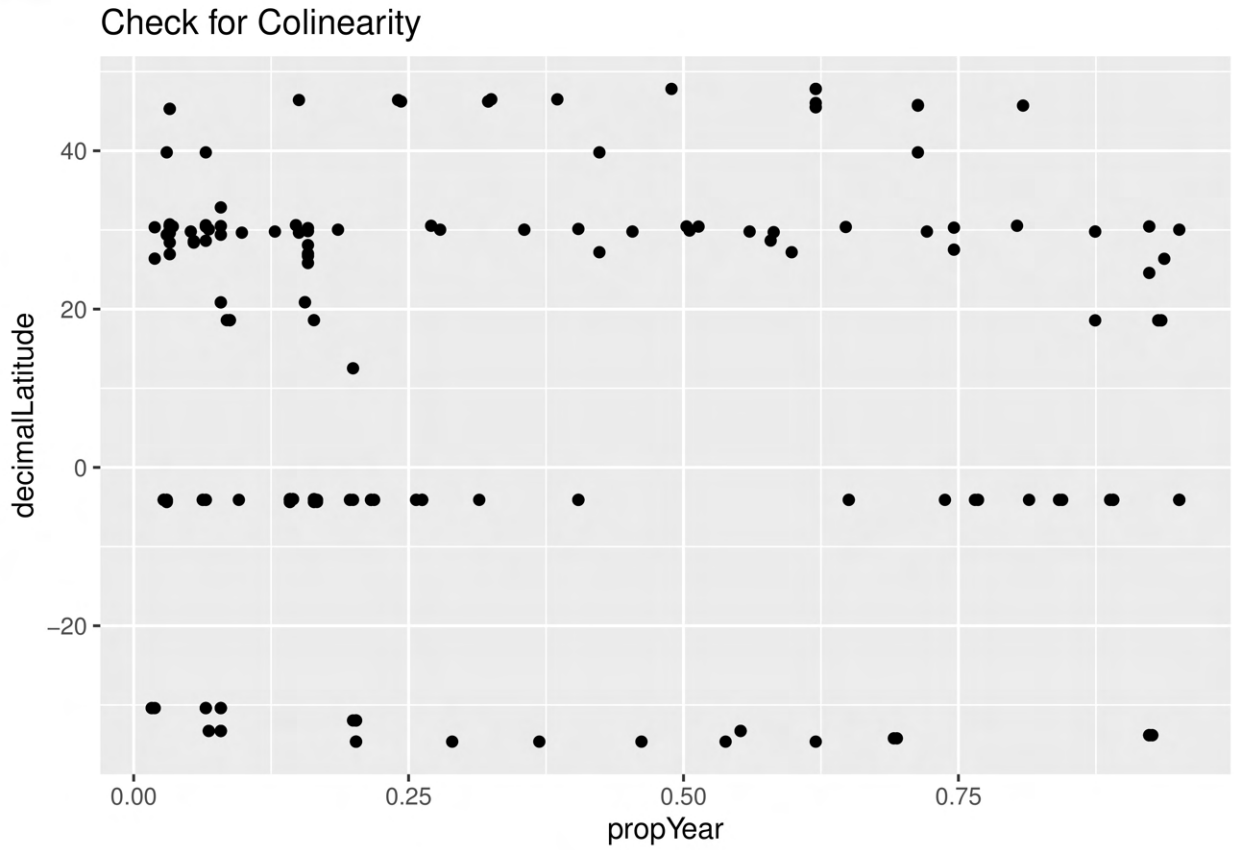
Table 27: Terms for Initial Model

term	estimate	p.value
(Intercept)	12.1315696	0.0000000
sexM	0.2094187	0.0000000
wgt	0.0497244	0.0000000
ageZ	0.3962320	0.5972905
species_idDM	21.1863051	0.0000000
species_idDO	20.5091757	0.0000000
species_idDS	31.4596477	0.0000000
species_idOL	6.4442566	0.0000000
species_idOT	6.4514445	0.0000000
species_idOX	6.1191195	0.0000000
species_idPB	12.0181841	0.0000000
species_idPE	6.5355170	0.0000000
species_idPF	2.4977812	0.0000000
species_idPH	11.9320586	0.0000000
species_idPI	8.8981272	0.0000000
species_idPL	6.4770678	0.0000000
species_idPM	6.7835086	0.0000000
species_idPP	8.2681699	0.0000000
species_idPX	5.4777113	0.0000346
species_idRF	4.2550877	0.0000000
species_idRM	3.1892921	0.0000000

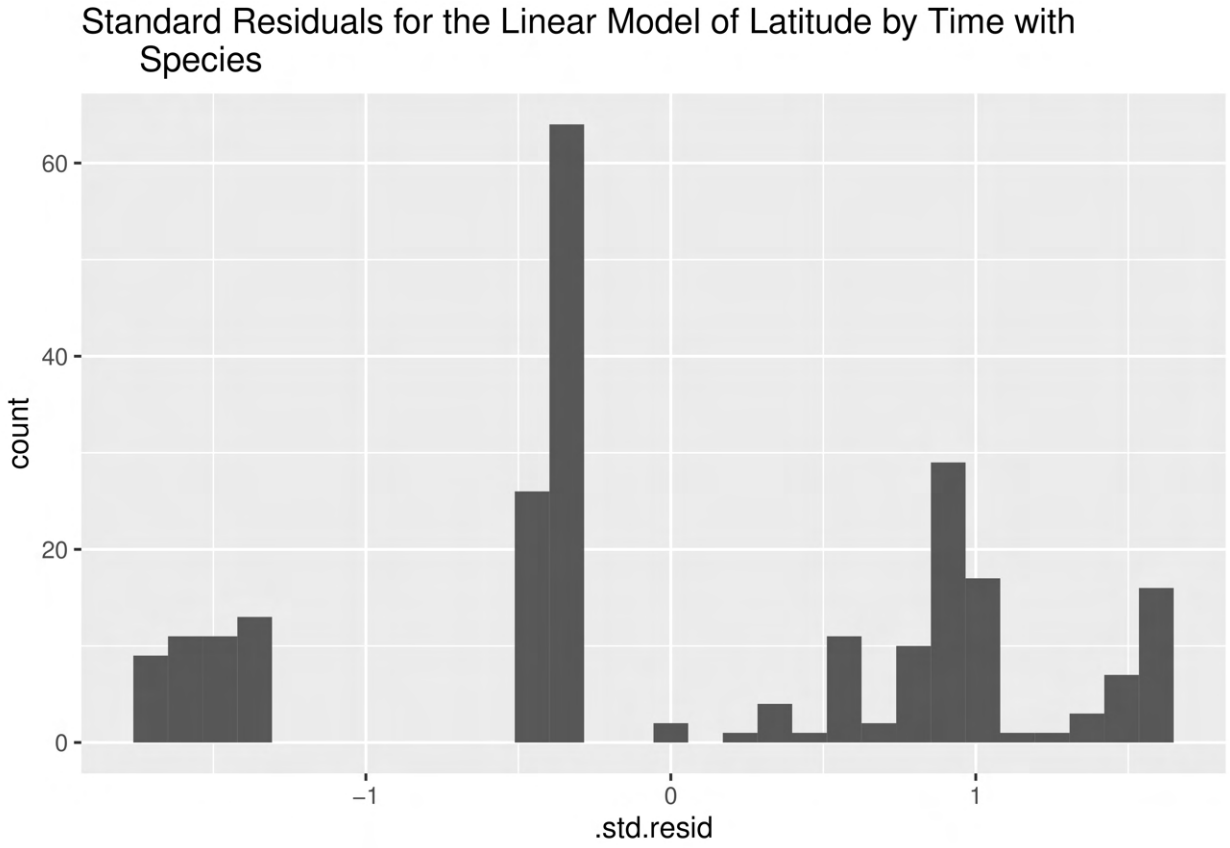
term	estimate	p.value
species_idRO	2.9182163	0.0002268
species_idSF	11.4464062	0.0000000
species_idSH	12.5721553	0.0000000
species_idSO	10.6498463	0.0000000

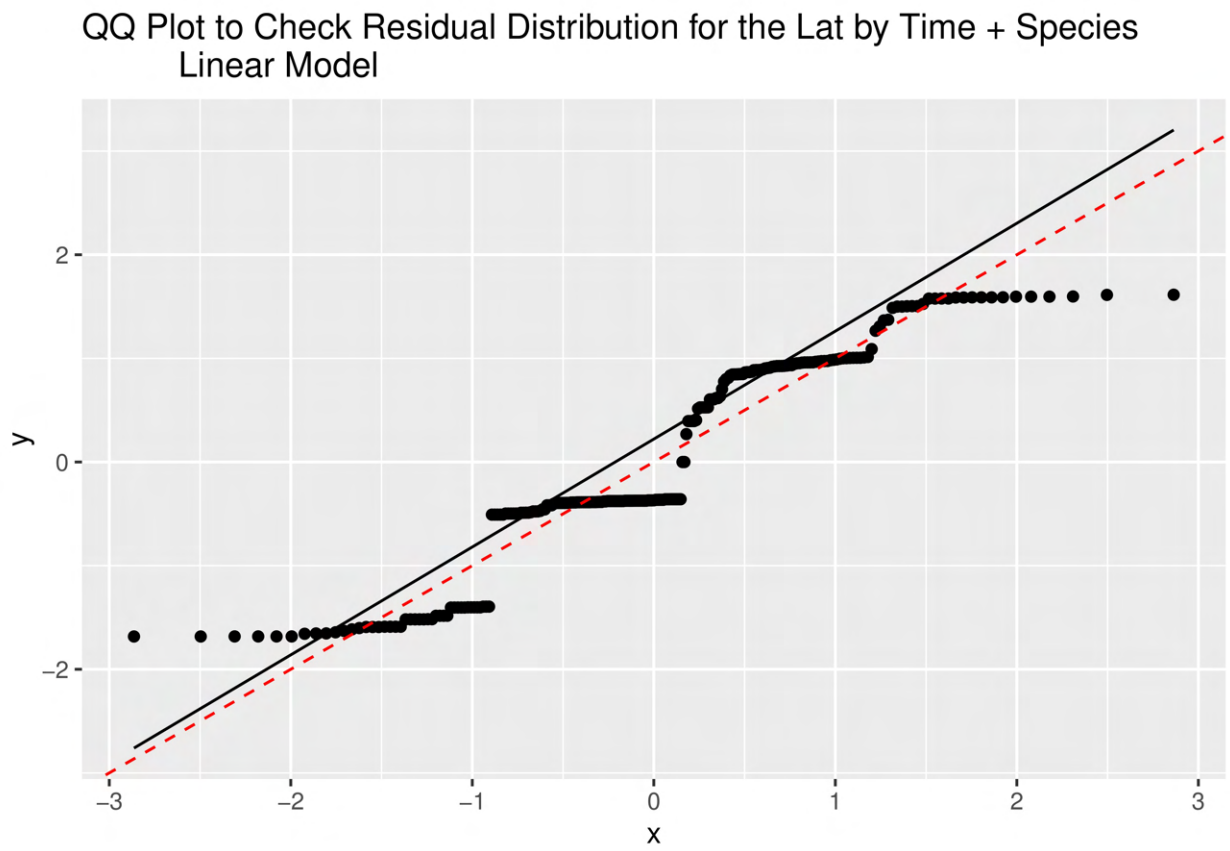
HYPOTHESIS IV

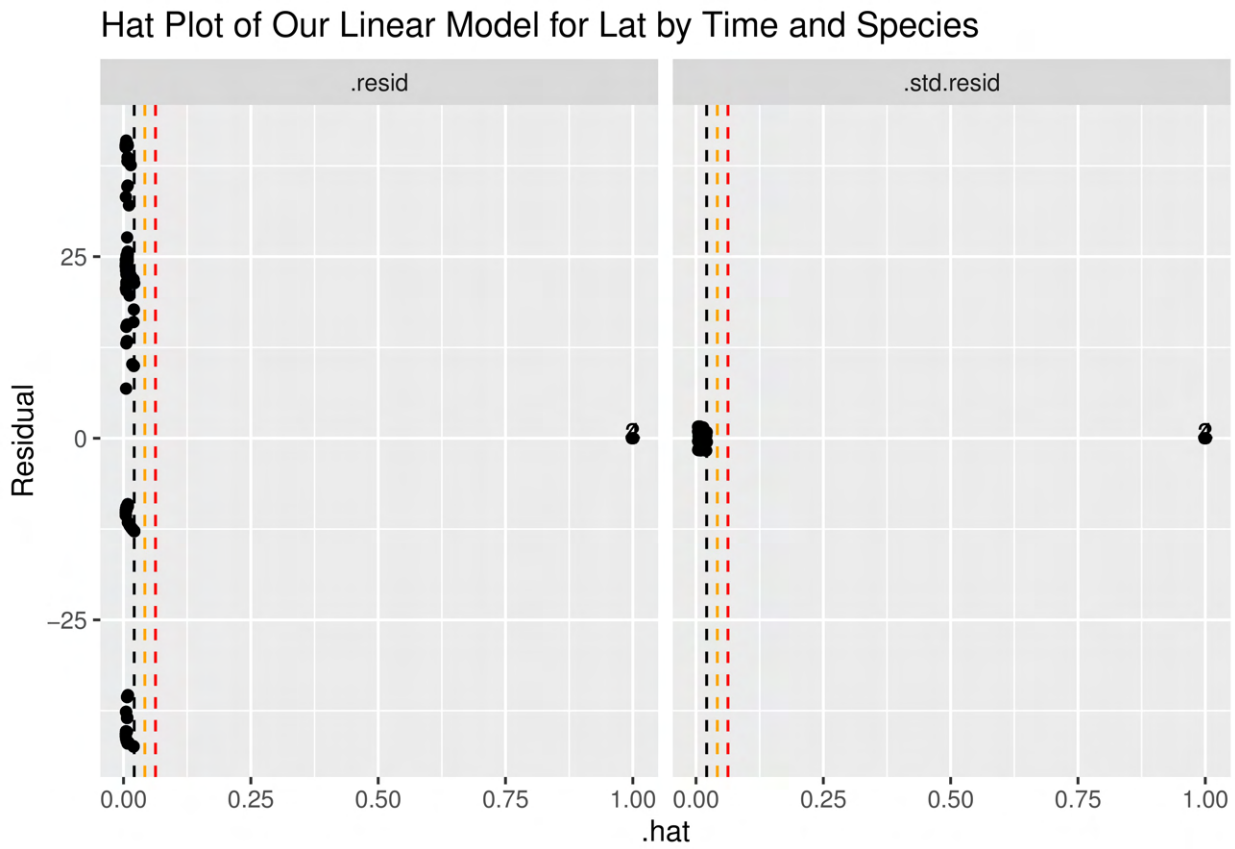
CHECK FOR COLINEARITY

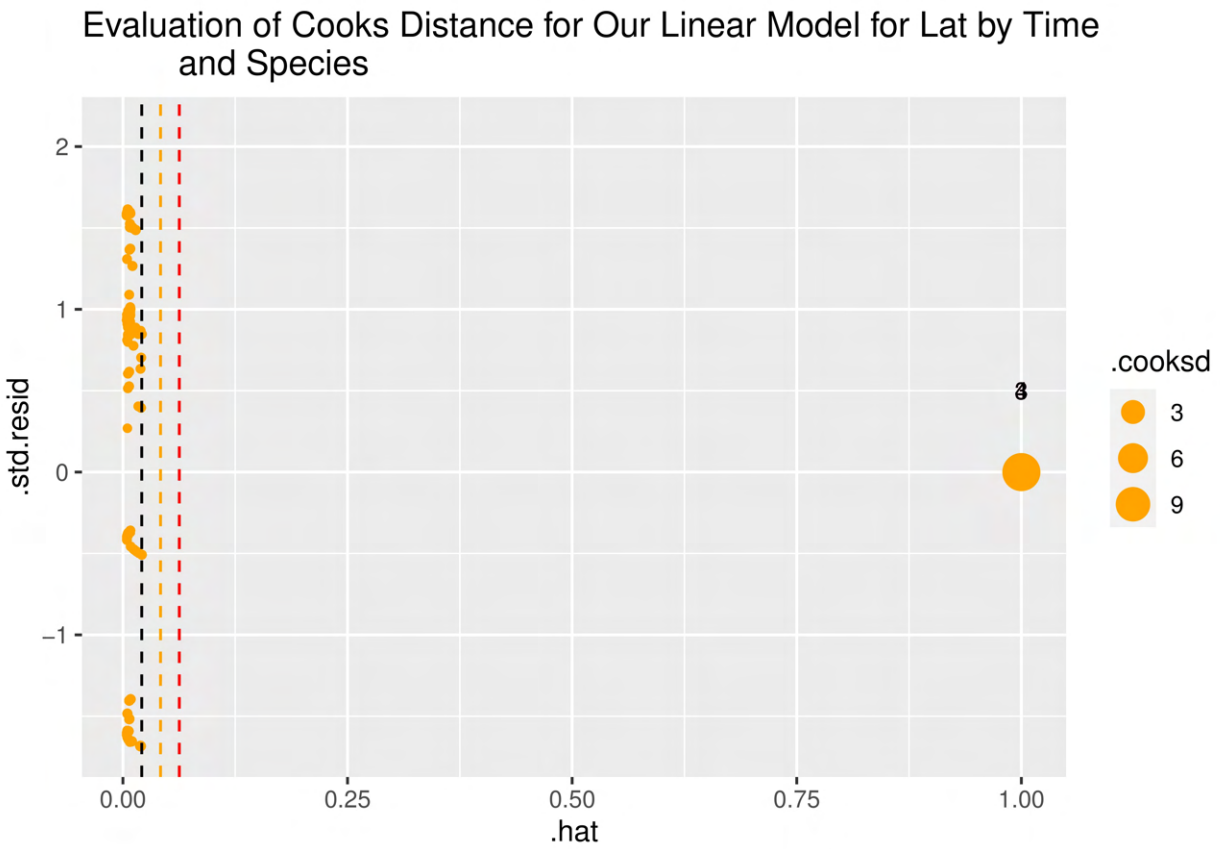


OUTLIER CHECK




QQ Plot to Check Residual Distribution for the Lat by Time + Species Linear Model
y
x
2
0
−2
−3
−2
−1
0
1
2
3


Hat Plot of Our Linear Model for Lat by Time and Species
.resid
.std.resid
Residual
25
0
−25
0.00
0.25
0.50
0.75
1.00
.hat


Evaluation of Cooks Distance for Our Linear Model for Lat by Time and Species
.std.resid
2
1
0
−1
0.00
0.25
0.50
0.75
1.00
.hat
.cooksd
3
6
9