ANALYSIS OF PALMER ARCHIPELAGO PENGUINS

Team HAT
Howard Mach, Andy Chandler, Thompson Pham

Prediction Model on Flipper Length

Data Description 2
- Introduction 2
- Variable Attributes 2

Data Analysis 3
- Data Cleaning 3
- Exploratory Analysis 4
- Model Construction & Fitting 6
- Residual Analysis 8
- Influential Points Analysis 11
- Data Transformation 13

Conclusion 15

Reflection 15

Appendix 16
- Citations 16
- Team Roles 17
- Code 17

Data Description

Introduction

In the vast landscape of ecological studies, understanding the nuances of species-specific traits remains a crucial endeavor. Among the array of captivating avian species, the charismatic nature of penguins, particularly the Palmer penguins (Pygoscelis papua), captivates researchers' curiosity. These remarkable creatures, residing in the subantarctic region, stand as emblematic symbols of resilience and adaptability in the face of evolving environmental conditions. At the heart of their biological characteristics lies the enigmatic feature of flipper length, a pivotal aspect influencing their movement, thermoregulation, and overall survival. With an inherent understanding that these dimensions hold valuable insights into their adaptation strategies, researchers have looked into unveiling the intricacies between flipper length and various morphological and categorical attributes.

This report embarks on a journey into the world of predictive modeling, seeking to understand the complex interplay between the given characteristics of Palmer penguins and their flipper length. Through the use of a comprehensive dataset comprising a myriad of features, including body mass, species, sex, culmen length, and depth. By the end, we aim to construct a robust predictive framework that demystifies the predictive potential embedded within these characteristics, offering a pathway to estimate flipper length with enhanced accuracy and reliability.

The dataset employed in this analysis, first discovered on Kaggle by Parul Pandey, encapsulates a trove of invaluable information initially published in a research study conducted by Gorman KB, Williams TD, and Fraser WR in 2014. It is composed of a collection of observations, showcasing 344 instances with 17 distinct variables, culminating in a repository of 5,848 data points. These observations detail the lives of penguins inhabiting the Antarctic Peninsula, sourced from samples obtained at the Palmer Station. This data forms an integral component of the Long Term Ecological Research Network in Antarctica, dedicated to studying the diverse penguin populations across three distinct islands situated on or near the Palmer Archipelago. We decided on the following variables as our first base model:

Variable Attributes

Response Variable: Flipper Length (mm)

Predictor Variables:

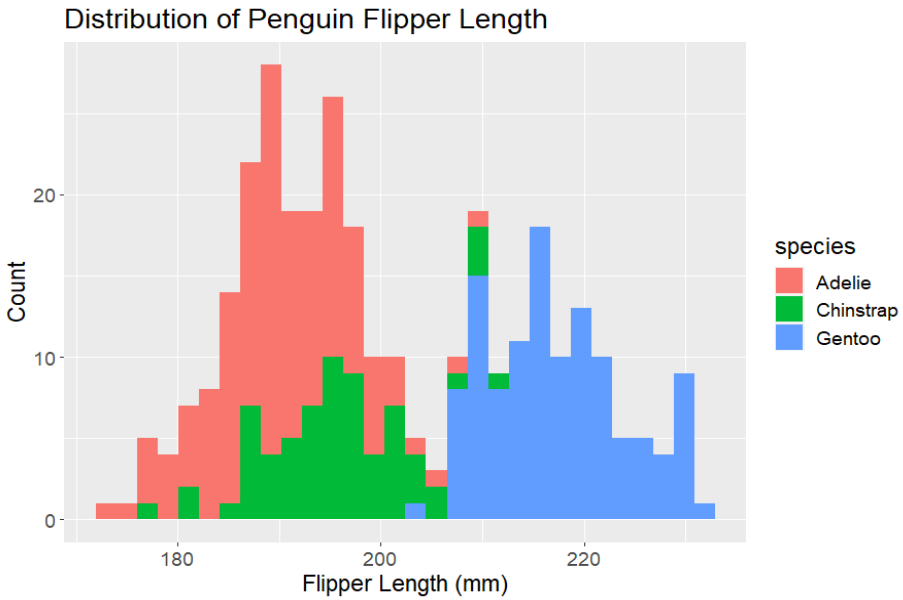
- Clutch Completion : Characteristic during the study denoting if the given nest observed is with a full clutch
- Culmen Length (mm) : Upper ridge of a bird's bill measured from the back to the front
- Culmen Depth (mm) : Upper ridge of a bird's bill measured from the top to the bottom
- Body Mass (g) : mass of the given penguin
- Delta 15 N (o/oo) : Stable isotope values of nitrogen, which can be found in organisms. For penguins it can be found in blood, feathers, and bone. Using measurements of the given data scientists can uncover foraging patterns, migratory behavior, and diet.
- Delta 13 C (o/oo) : Similar to Delta 15 N, but for carbon.
- Sex : Gender of the penguin
- Species : Adélie, Chinstrap, and Gentoo for our given data set
- Island : Islands penguins are found within the proximity to Palmer Station
- Date Egg : We first thought this name entailed the date an egg was hatched, and hoped to use it as a form of measurement for age. However we found it to be a misunderstanding on our part, it turned out to be the date each penguin was discovered and recorded in the study

Data Analysis

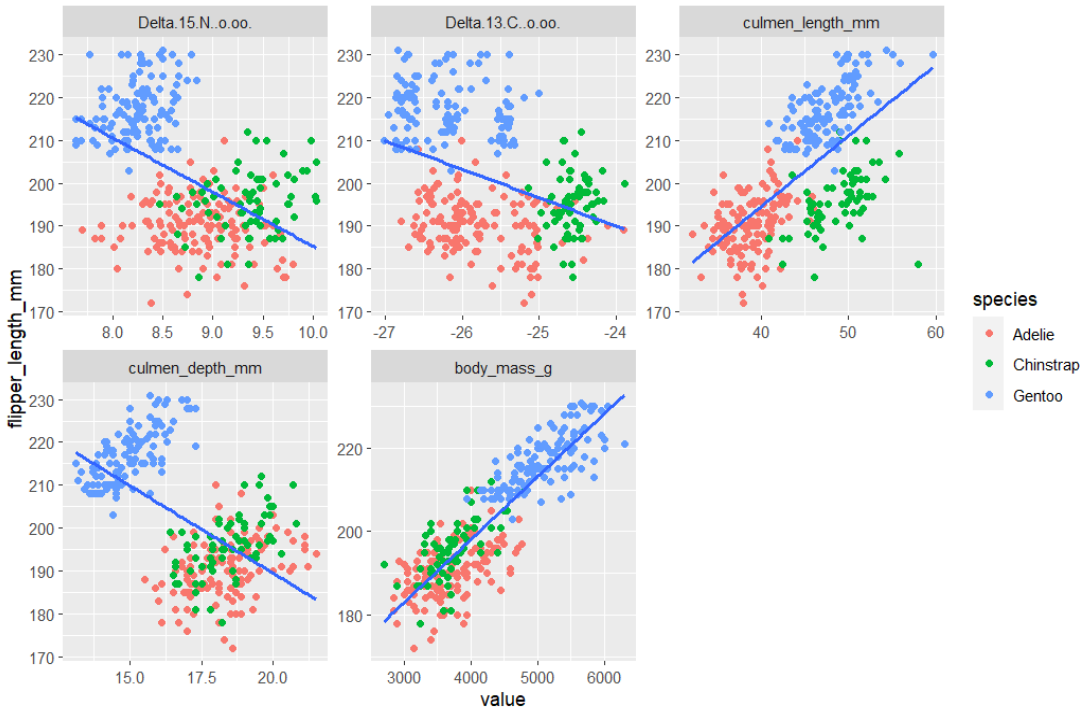
Data Cleaning

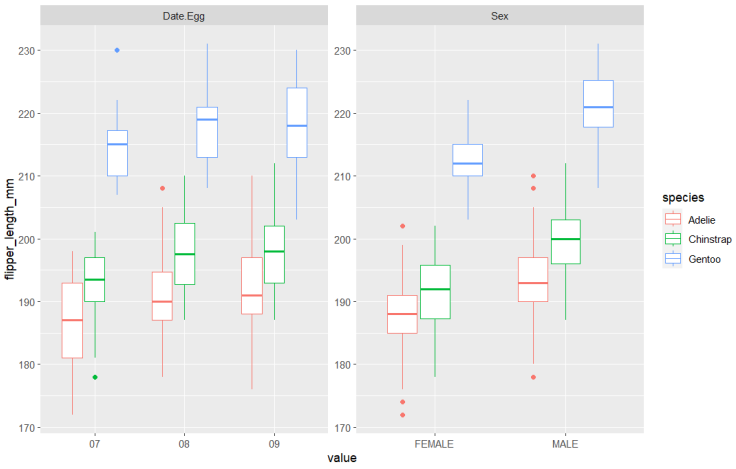
During the initial preprocessing stage, the data set we found on Kaggle was divided into two tables. So we first merged the two together and then refined it down. After the merge, we found the two data sets consisted of instances of the same variable and values, but labeled with marginally different names (Ex: Culmen.Length..mm. & culmen_length_mm). So we removed each “duplicate” column and also removed 5 variables that weren’t indicative to our goal, removing a total of 13 columns. Afterwards, we removed one row (Penguin) that gave an error value of “.” for gender and reformed the values under the egg date variable to show it by years.

Exploratory Analysis



First we aimed to build an overall distribution graph for our response variable, flipper length. The overall range of the given graph above is from 172 - 231 (mm), and within that given range we can get subranges for each respective species. Adelie varies from 172 - 210 (mm) with Chinstrap largely overlapping between 178 - 212 (mm). And Gentoo being a general outlier of the given range 203 - 231 (mm). We can see a bimodal distribution based on the graph above showcasing Adelie & Chinstrap on the smaller side as compared to Gentoo. After looking at distribution in relation to species, we turned towards predictor variables to see if there was any correlation they may have towards flipper length.





From our given predictor variables, we saw there was a mix of discrete and continuous variables together and decided the best course of action was to separate the two and build plots for their respective counterparts.

At a quick glance for continuous variables, there appears to be a strong linear relationship for body mass, culmen depth and length. With it waning a bit for the two given delta values; body mass and culmen length being a positive trend and the rest are negative. Similar to the overall distribution map above we can see some overlap between Adelie and Chinstrap, indicating some form of correlation between the two. For the discrete variables we can't really conclude a linear relationship, but we do see a general correlation between Adelie and Chinstrap as shown by the other two plots.

Model Construction & Fitting

In our regression analysis we sought to determine which of our remaining variables were viable to use for our model to predict flipper length, in millimeters, by using the backward selection approach.

```
> summary(fullModel)

Call:
lm(formula = flipper_length_mm ~ Delta.15.N..o.oo. + Delta.13.C..o.oo. +
    culmen_depth_mm + culmen_length_mm + body_mass_g, data = PenguinOmit)

Residuals:
     Min       1Q   Median       3Q      Max
-22.3776  -3.3046   0.0945   3.7058  14.0162

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       73.0454063 18.2330264   4.006 7.69e-05 ***
Delta.15.N..o.oo.  0.5456396  0.8325680   0.655    0.513
Delta.13.C..o.oo. -2.7833034  0.5371210  -5.182 3.91e-07 ***
culmen_depth_mm   -1.3509443  0.2039209  -6.625 1.48e-10 ***
culmen_length_mm   0.7809756  0.0855588   9.128  < 2e-16 ***
body_mass_g        0.0096432  0.0006524  14.781  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.499 on 318 degrees of freedom
Multiple R-squared:  0.8472,    Adjusted R-squared:  0.8448
F-statistic: 352.7 on 5 and 318 DF,  p-value: < 2.2e-16
```

From this output we can see that for the most part our regressors are quite significant with incredibly small p-values. The only standout is Delta 15 with a p-value greater than 0.5. To explore why Delta 13 is significant where Delta 15 is not, and what as well as where they came from, we decided to do some additional research from the study that is the source of our dataset. Both Delta 13 and Delta 15 are isotopes of carbon and nitrogen, respectively, that were observed within the tested penguins. These most likely come from differences in diet of the penguins, more specifically the krill that is closest to the island they inhabit. While both isotopes were observed within a given penguin, only Delta 13 showed any statistical relevance to the measurements of the penguin, with Delta 15 not having a notable positive or negative correlation. Given this information, we constructed a reduced model.

```
Call:
lm(formula = flipper_length_mm ~ Delta.13.C..o.oo. + culmen_depth_mm +
    culmen_length_mm + body_mass_g, data = PenguinOmit_reduced)

Residuals:
     Min       1Q   Median       3Q      Max
-14.9469  -3.2755  -0.1467   3.5655  14.0716

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       74.9158237 15.3727667   4.873 1.74e-06 ***
Delta.13.C..o.oo. -2.8069258  0.5091072  -5.513 7.29e-08 ***
culmen_depth_mm   -1.2942930  0.1806431  -7.165 5.45e-12 ***
culmen_length_mm   0.8636770  0.0824517  10.475  < 2e-16 ***
body_mass_g        0.0091095  0.0005937  15.343  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.348 on 318 degrees of freedom
Multiple R-squared:  0.8545,    Adjusted R-squared:  0.8527
F-statistic:   467 on 4 and 318 DF,  p-value: < 2.2e-16
```

With our reduced model we can observe an increase in our adjusted R^2 value from the previous 0.8448 to 0.8527. In addition, all of the other regressors still fall under our significance level so no further alterations were necessary. We then moved forward to perform an ANOVA analysis to compare our two models.

```
> anova(reducedModel, fullModel)
Analysis of Variance Table

Model 1: flipper_length_mm ~ Delta.13.C..o.oo. + culmen_depth_mm + culmen_length_mm +
    body_mass_g
Model 2: flipper_length_mm ~ Delta.15.N..o.oo. + Delta.13.C..o.oo. + culmen_depth_mm +
    culmen_length_mm + body_mass_g
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    318 9095.2
2    317 9089.1  1    6.0961 0.2126  0.645
```

From this we can see that our reduced model is the better fit for our data with a p-value of 0.645. To ensure that our reduced model is in fact the best fit we then moved on to perform an analysis for any potential multicollinearity.

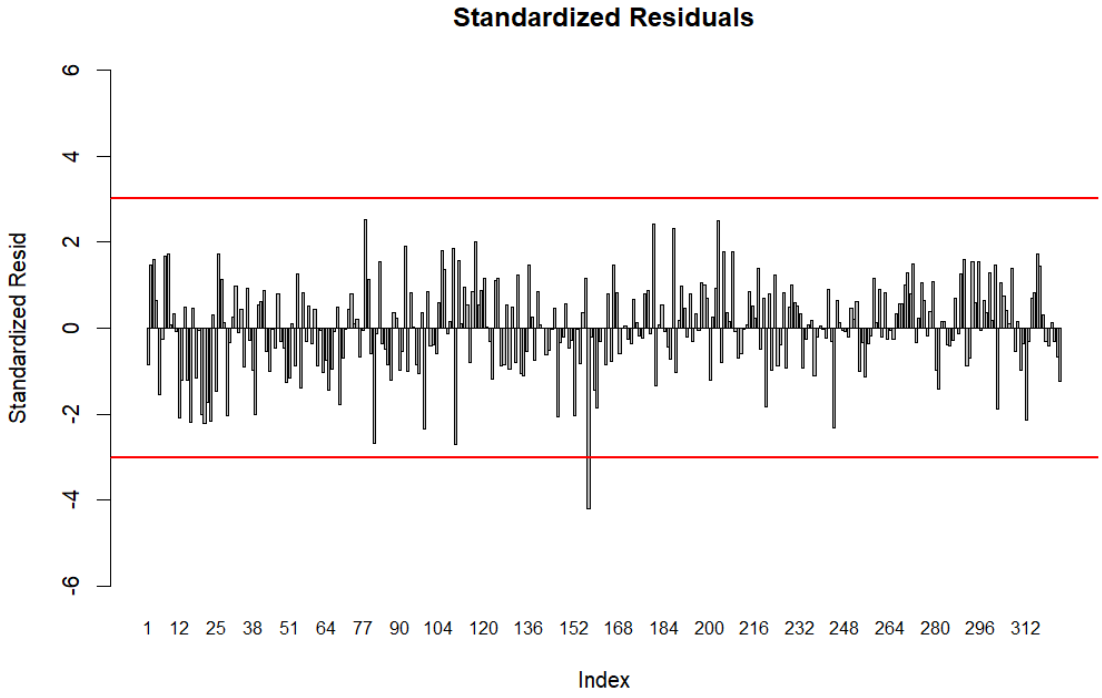
Variable	Delta.13.C..o.oo.	culmen_depth_mm	culmen_length_mm	body_mass_g
VIF	1.819519	1.425882	2.259734	2.603961

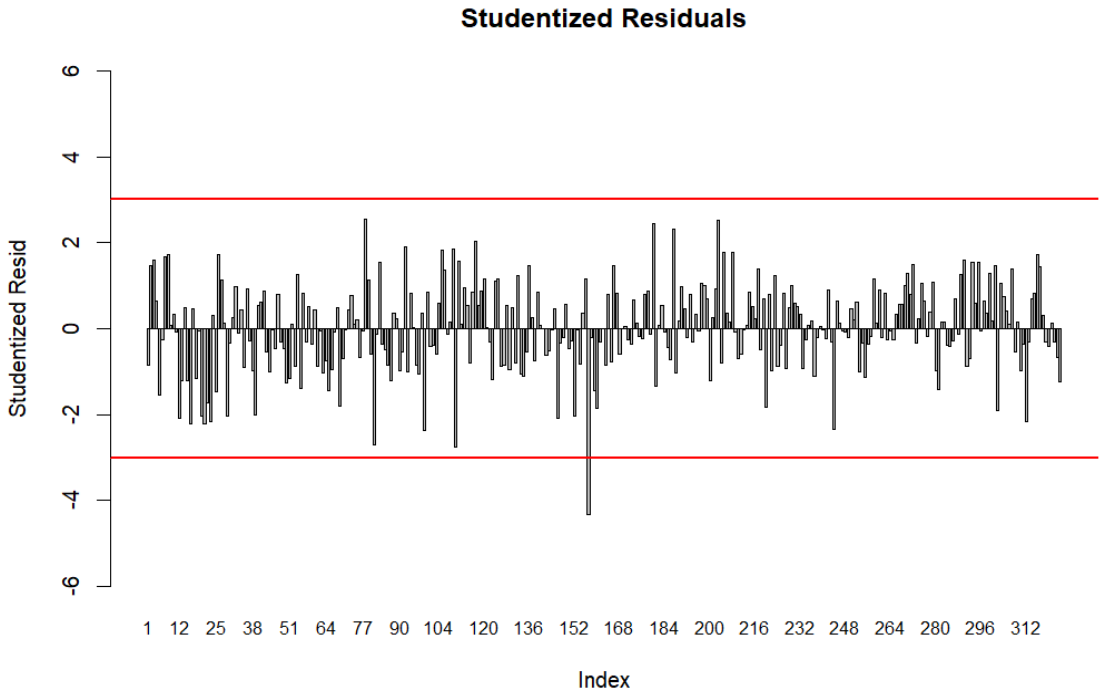
Since all the VIF values are significantly below 10, then there is no evidence that our model has multicollinearity within it. This was not a shocking revelation for us as with

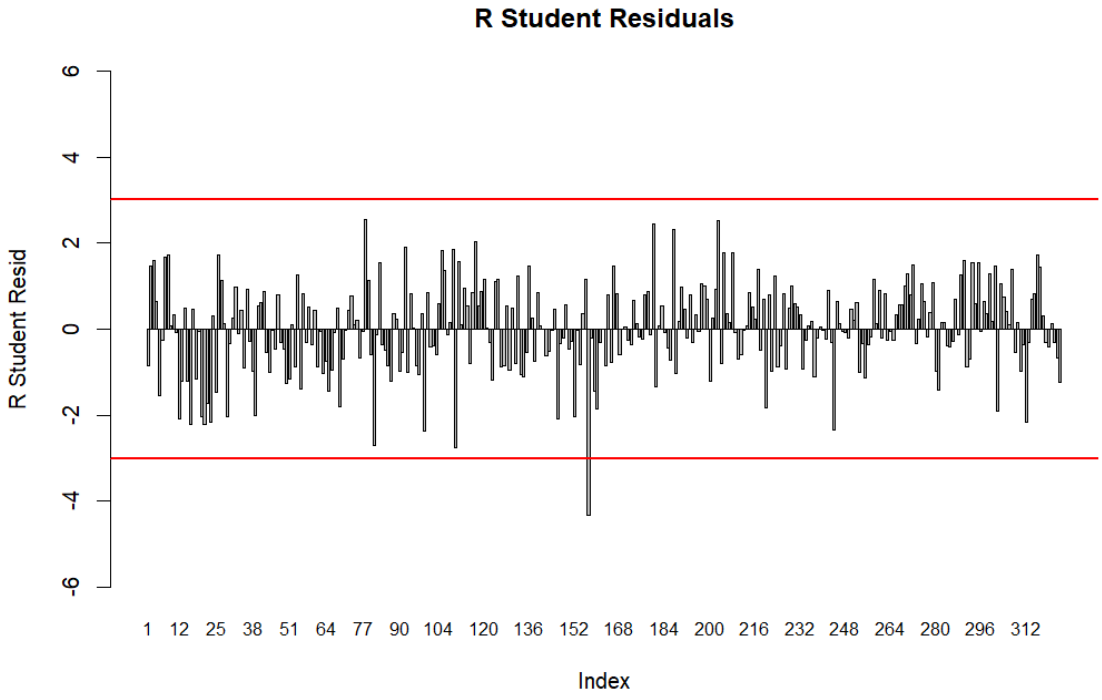
the various measurements of the penguins, in addition to the different species observed, are not typically directly indicative of a different measurement since with wildlife growth there are many factors that affect those measurements that were not included within our data set such as genetic disposition or any influences from nature. With this process complete we concluded that our reduced model is the best fit for our analysis, and that no further alterations would be required.

Residual Analysis

After creating our model, we conducted a residual analysis to identify any possible outliers and influential points within the observations. This analysis would also allow us to determine overall model adequacy and whether we need to look into performing a transformation or any other analyses. Our residual analysis consisted of plotting the standardized, studentized, and r-student residual plots. We used cutoff values of 3 and -3 as they are the standard values to determine the spread of the data.

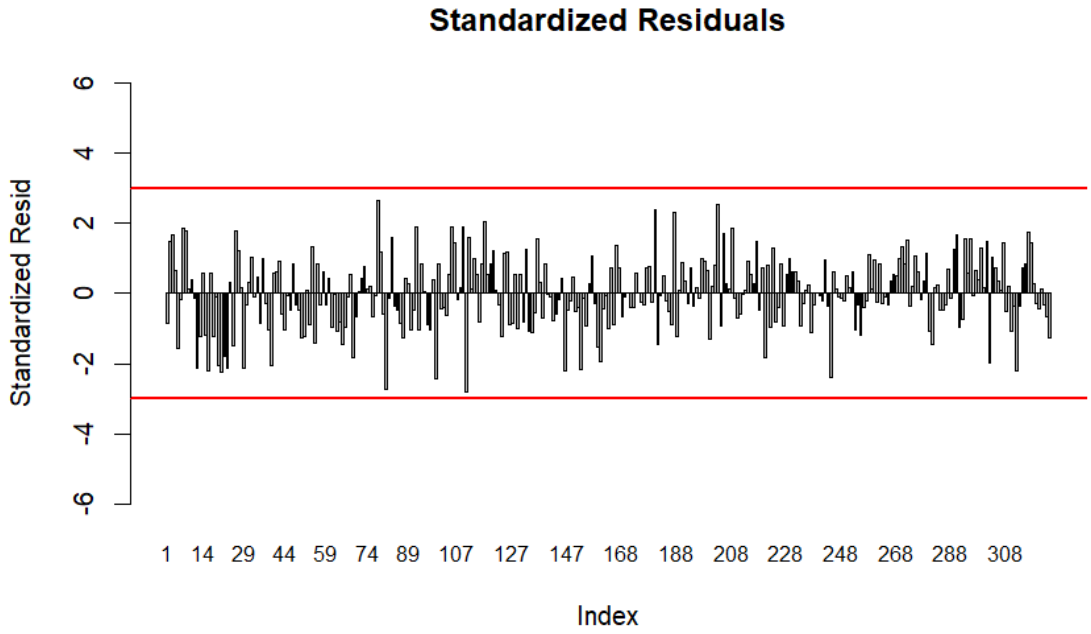


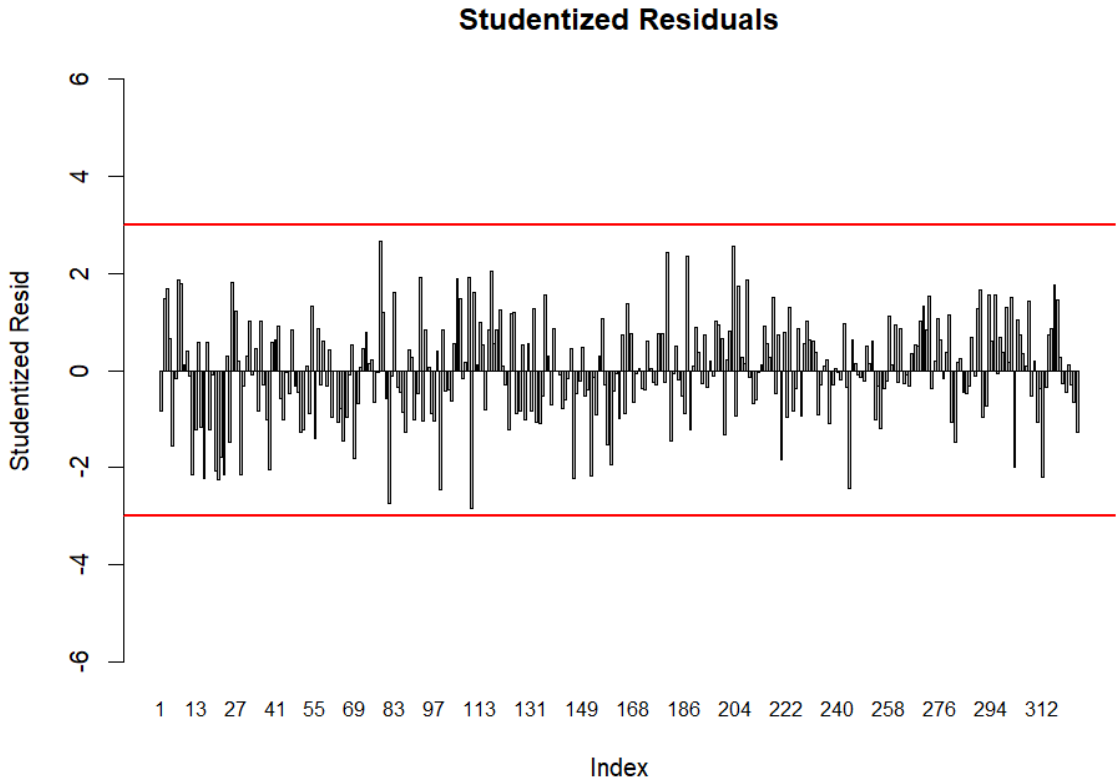


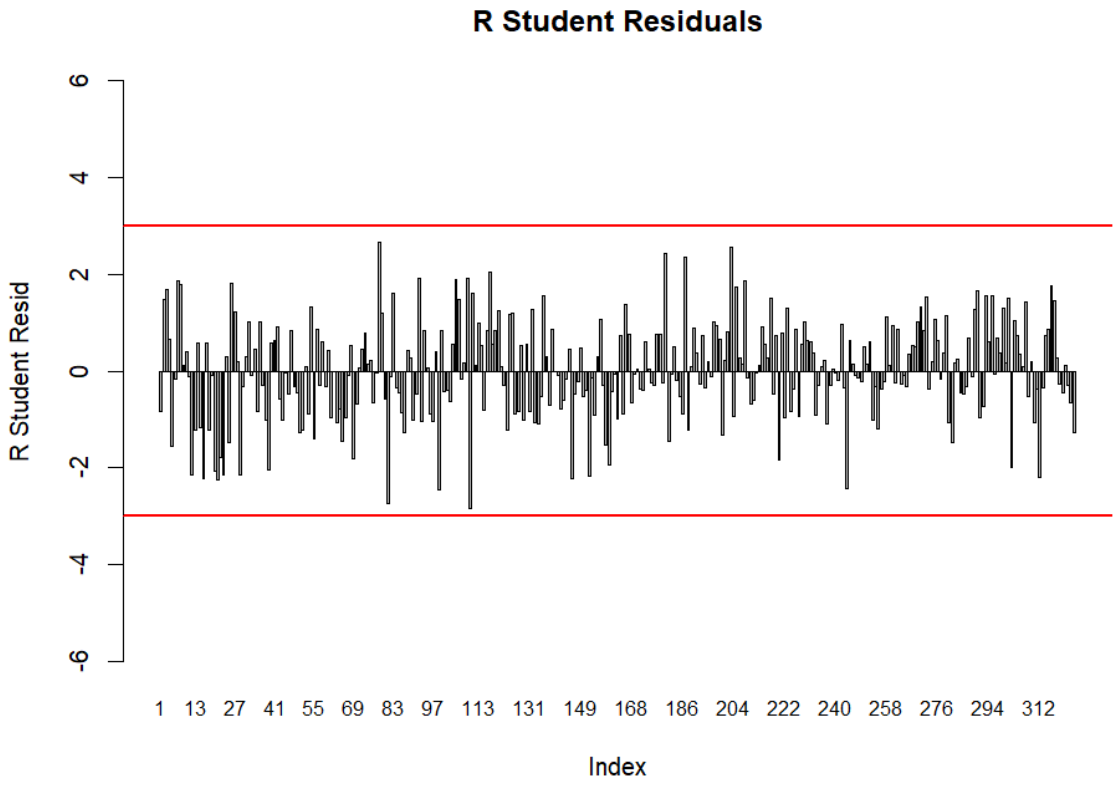


We noticed in the model that one of the residuals lies beyond the cutoff value. That specific observation, 157, needs to be looked at as a possible outlier. Looking at the other observations, we see that they are within the cutoff values. To address this, we

adjusted the model so that it does not include that specific observation. After doing that, we got the residual plots below.



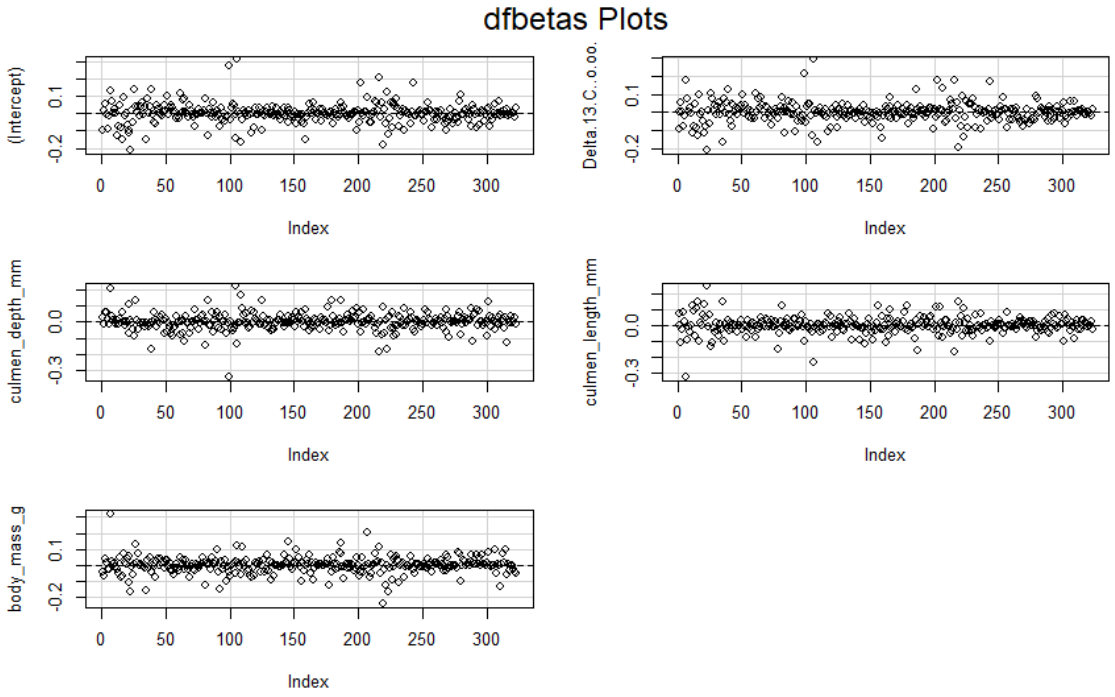


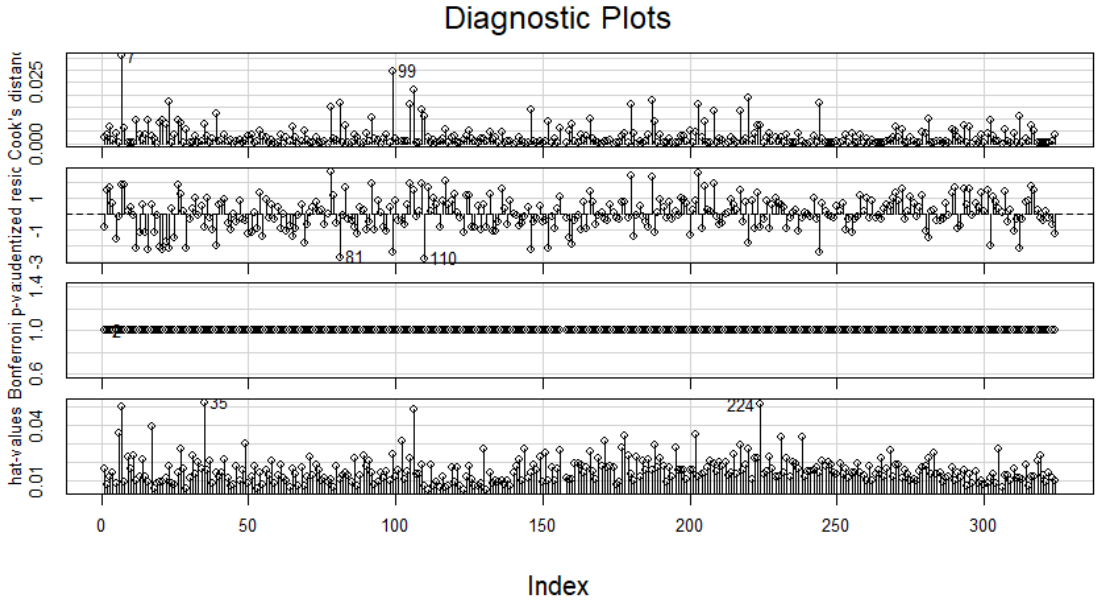


Now, every residual in the new plots resides within the cutoff values, meaning that we have no possible outliers. This means that our data is good as is. However, we need to do some more investigation to determine if there are some influential points or if there is a need to transform the data.

Influential Points Analysis

We now move on to our influential analysis. To perform this thoroughly, we use all measures from Cook's Distance, DFBETAS (one for each predictor), DFFITS, COVRATIO, and hat values.





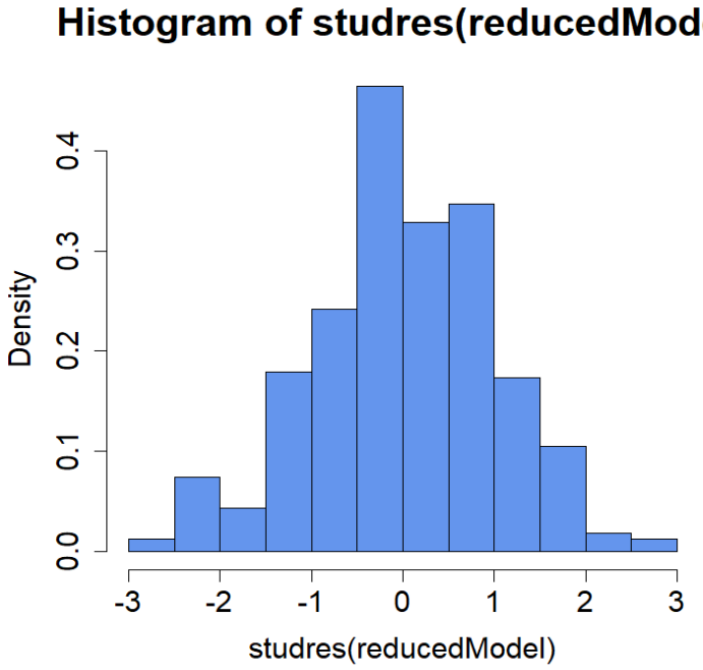
We observed that there are 19 influential points, which make up 5.88% of the dataset. This high percentage implies that some of the observations will have a dramatic impact on our model statistics despite not being outliers. 4 of the 19 points are leverage points, some of them are both influential and leverage points at the same time.

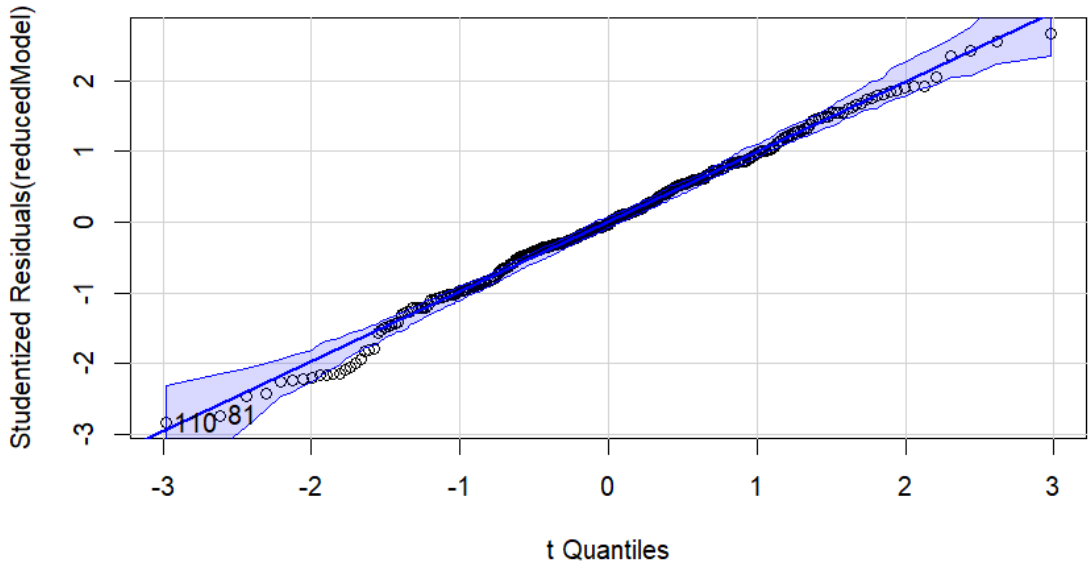
Type of measures	Hat values	DFBETAS	DFFITS	COVRATIO	Cook's D
Number of points	4	9	2	17	0

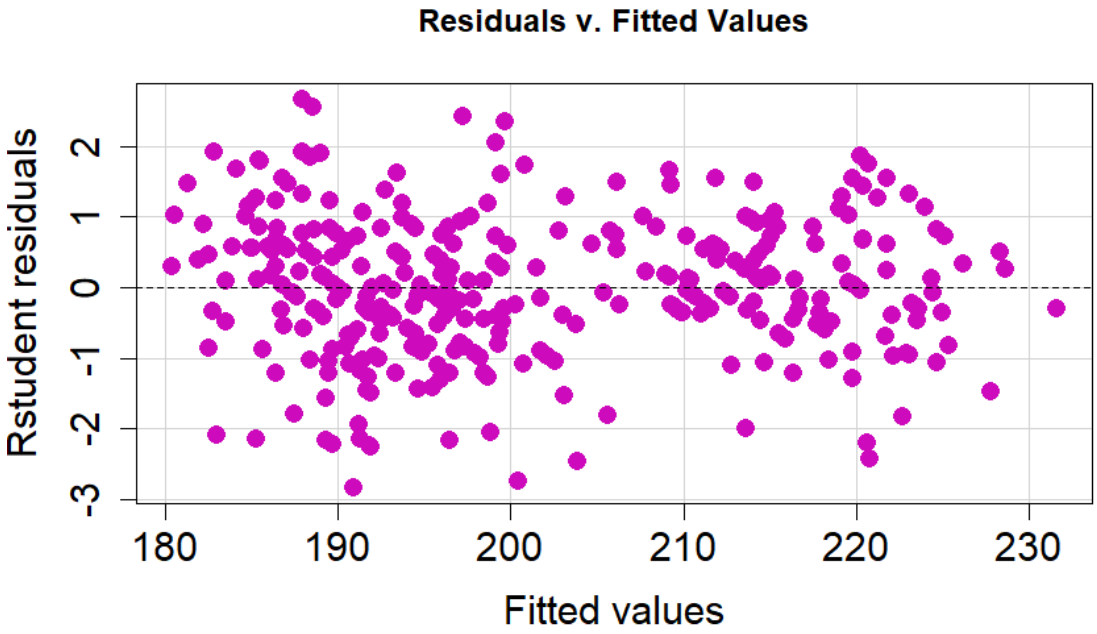
To further understand this result, we take a closer look at some observations, specifically numbers 7, 99, and 106. These data points were of the Adelie species, meaning that species of penguin have some interesting characteristics. Therefore, by influential analysis, we are not only able to investigate any possible influential points, but also we can understand more about a particular species of penguin.

Data Transformation

We decided to view the histogram and QQ plot of residuals to check whether our normality assumption was correct or not. Before any transformation, we observed that the histogram was not skewed. In addition, the points on the QQ plot lay mostly on a straight diagonal line, indicating that our normality assumption was correct. For the residuals versus fitted values plot, the points ranged from -3 to 3 but converged around the zero line. This indicates that our model may not require a transformation.

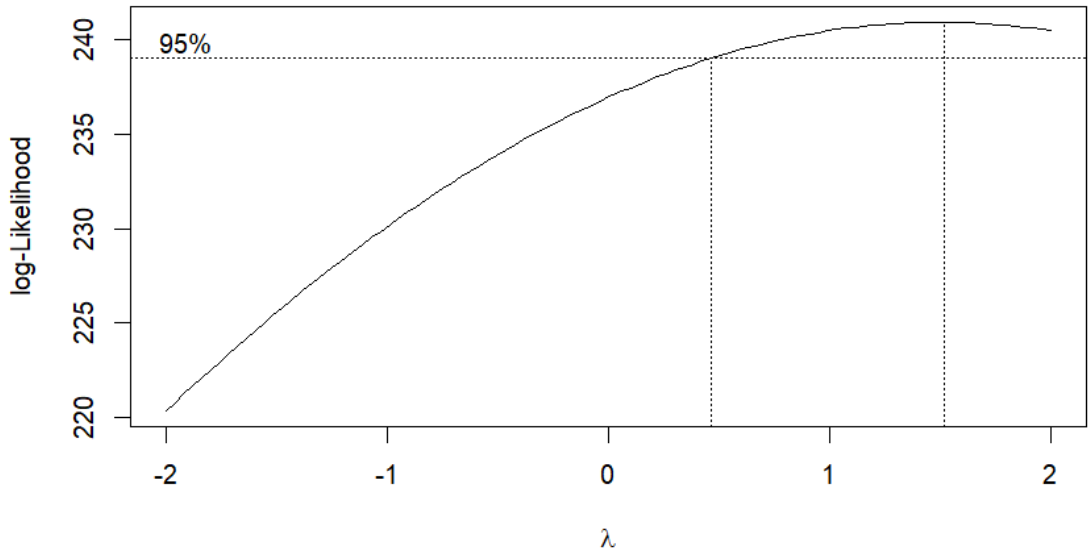






However, we decided to use the Box-Cox method to prove that we do not need to transform the data. It is important to note that for the Box-Cox method to work, the response variable must be strictly positive. Our response variable, flipper length, has all positive values in our dataset, so no further action is needed. After running the test, we obtained a lambda value of 1.51. However, since the lambda value of 1 is in the 95%

confidence range of the Box-Cox plot, it is deemed unnecessary to perform a transformation on our data.



Conclusion

Our analysis reveals that the flipper length is determined by the following model:

Y = 74.9158237 - 2.8069258*Delta_13_C - 1.2942930*Culmen_Depth_mm + 0.8636770*Culmen_Length_mm + 0.0091095*Body_Mass_g

Reflection

This project was great in allowing us to apply the skills and analyses that we learned in this class to a dataset. By choosing to analyze the penguin dataset, we could determine how the other factors influence the flipper length of a penguin. Some aspects of our analysis went well. Notably, we were able to find a dataset that was easy to clean and analyze. After, we were able to create a reduced model and determine whether it was an adequate model or required a transformation. We were able to follow the steps of residual analysis and create a model for the data.

On the other hand, some things made our analysis more difficult. This dataset was created by taking data published by the research study. This helped to ensure the

accuracy of the data, but unfortunately, they do not release the methods that they use to come up with some of these values. This sometimes made it more difficult to interpret results since we were not sure how some of these numbers were calculated. Also, in the beginning, it was difficult to coordinate our code since we didn't have a way to edit our code at once. Instead, we used this report as a way to put our code together so that each of us could run it on our software.

Another part of the project that gave us a little bit more to think about was the outliers in the data. We had come up with a model that represented the data, but when it came to do our residual analysis, it showed an outlier in each of the plots. At first, we didn't know what to do with it. After some time, we decided to remove the outlier since the rest of the plot looked fine. As a result, we had to make a new dataset with the outlier point removed and then construct a new model based on that dataset. This part is discussed more in the residual analysis section.

Once we had our new model, we were able to perform further analysis of our data, including transformation. In the future, we hope to look at a variety of other species and include a more varying list of characteristics in our regressor variables to provide a more comprehensive analysis. We could also look at other datasets with the following information above.

Appendix

Citations

Palmer Station Antarctica LTER and K. Gorman. 2020. Structural size measurements and isotopic signatures of foraging among adult male and female Adélie penguins (Pygoscelis adeliae) nesting along the Palmer Archipelago near Palmer Station, 2007-2009 ver 5. Environmental Data Initiative. https://doi.org/10.6073/pasta/98b16d7d563f265cb52372c8ca99e60f

Palmer Station Antarctica LTER and K. Gorman. 2020. Structural size measurements and isotopic signatures of foraging among adult male and female Gentoo penguin (Pygoscelis papua) nesting along the Palmer Archipelago near Palmer Station, 2007-2009 ver 5. Environmental Data Initiative. https://doi.org/10.6073/pasta/7fca67fb28d56ee2ffa3d9370ebda689

Palmer Station Antarctica LTER and K. Gorman. 2020. Structural size measurements and isotopic signatures of foraging among adult male and female Chinstrap penguin (Pygoscelis antarctica) nesting along the Palmer Archipelago near Palmer Station,

2007-2009 ver 6. Environmental Data Initiative.
https://doi.org/10.6073/pasta/c14dfcfada8ea13a17536e73eb6fbe9e

Gorman KB, Williams TD, Fraser WR (2014) Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins (Genus Pygoscelis). PLoS ONE 9(3): e90081. doi:10.1371/journal.pone.0090081.
https://www.kaggle.com/datasets/parulpandey/palmer-archipelago-antarctica-penguin-data/data

https://allisonhorst.github.io/palmerpenguins/reference/penguins_raw.html

Team Roles

Howard Mach - Residual Analysis, Influential Points Analysis, Histogram and QQ Plot for Transformation, Box-Cox Plot, Conclusion

Andrew Chandler - Model Fitting, Model Reduction, ANOVA Analysis, Variance Inflation Factors

Thompson Pham - Introduction, Variable Attributes, Data Cleaning, Exploratory Analysis

Code

Data Cleaning

```
# importing libraries
library(dplyr)
library(tibble)
library(ggplot2)
library(MASS)
library(reshape2)
library(car)

# loading and cleaning data
basePenguin_lter <- read.csv("penguins_lter.csv")
basePenguin_size <- read.csv("penguins_size.csv")
PenguinDF <- cbind(basePenguin_lter, basePenguin_size)

PenguinOmit <- na.omit(PenguinDF)
```

```
PenguinOmit <- subset(PenguinOmit, select = -c(1, 2, 3, 4, 5, 6, 7, 10, 11, 12, 13, 17,
24))
PenguinOmit <- dplyr::filter(PenguinOmit, Sex %in% c('MALE', 'FEMALE'))
PenguinOmit$Date.Egg <- format(as.Date(PenguinOmit$Date.Egg,
format="%m/%d/%Y"),"%Y")
PenguinOmit$Date.Egg <- substring(PenguinOmit$Date.Egg, 3)

# characteristics of numeric data
summary(PenguinOmit$culmen_length_mm)
summary(PenguinOmit$culmen_depth_mm)
summary(PenguinOmit$flipper_length_mm)
summary(PenguinOmit$body_mass_g)

# more characteristics of numeric data
var(PenguinOmit$culmen_length_mm)
range(PenguinOmit$culmen_length_mm)

var(PenguinOmit$flipper_length_mm)
range(PenguinOmit$flipper_length_mm)

var(PenguinOmit$culmen_depth_mm)
range(PenguinOmit$culmen_depth_mm)

var(PenguinOmit$body_mass_g)
range(PenguinOmit$body_mass_g)
```

Exploratory Data Analysis

```
# distribution of numeric variables

ggplot(PenguinOmit, aes(x = flipper_length_mm, fill = species)) +
  geom_histogram()+
  ggtitle("Distribution of Penguin Flipper Length") +
  theme(text = element_text(size = 14)) +
  labs(x = "Flipper Length (mm)", y = "Count")

range_adelie <- range(PenguinOmit$flipper_length_mm[PenguinOmit$species ==
"Adelie"])
print(range_adelie)
```

```
range_chinstrap <- range(PenguinOmit$flipper_length_mm[PenguinOmit$species ==
"Chinstrap"])
print(range_chinstrap)
range_gentoo <- range(PenguinOmit$flipper_length_mm[PenguinOmit$species ==
"Gentoo"])
print(range_gentoo)

PenguinOmit %>%
  group_by(species) %>%
  summarise(flipper_length_mm = paste(min(flipper_length_mm), "-",
max(flipper_length_mm)))

# formatting egg data
melt <- dplyr::select(PenguinOmit, -Clutch.Completion)

meltcont <- dplyr::select(melt, -island, -Sex, -Date.Egg)
melt$species <- as.factor(melt$species)
melt$Sex <- as.factor(melt$Sex)

# flipper length
flipper_lengthdfl <- melt(data = meltcont, id = c("species", "flipper_length_mm"))
```

Variable Selection

```
# choosing good response variable
ggplot(data = flipper_lengthdfl, aes(x = value, y = flipper_length_mm)) +
  geom_point(aes(color = variable)) + facet_wrap(~variable, scales = "free") +
  geom_smooth(method = "lm", se = FALSE)

ggplot(data = flipper_lengthdfl, aes(x = value, y = flipper_length_mm)) +
  geom_point(aes(color = species)) + facet_wrap(~variable, scales = "free") +
  geom_smooth(method = "lm", se = FALSE)

# making boxplot of flipper length v. discrete variables
meltdisc <- dplyr::select(melt, species, Date.Egg, Sex, flipper_length_mm)

flipper_lengthdfl2 <- melt(data = meltdisc, id = c("species", "flipper_length_mm"))
ggplot(flipper_lengthdfl2, aes(x = value, y = flipper_length_mm)) +
  geom_boxplot(aes(color = variable)) +
  facet_wrap(~variable, scales = "free")
```

```
ggplot(flipper_lengthdfl2, aes(x = value, y = flipper_length_mm)) +
  geom_boxplot(aes(color = species)) +
  facet_wrap(~variable, scales = "free")
```

Model Construction

```
PenguinOmit_reduced <- PenguinOmit[-c(157),]

# constructing full linear model
fullModel <- lm(data = PenguinOmit_reduced, flipper_length_mm ~ Delta.15.N..o.oo. +
Delta.13.C..o.oo. + culmen_depth_mm + culmen_length_mm + body_mass_g)

summary(fullModel)
# Remove variables with p-value > 0.05
fullModel <- lm(data = PenguinOmit_reduced, flipper_length_mm ~ Delta.13.C..o.oo. +
culmen_depth_mm + culmen_length_mm + body_mass_g)
```

Variance Inflation Factors

```
# variance inflation factors of model
vif(fullModel)
```

ANOVA

```
# constructing reduced model (removing variables with p-value > 0.05)
reducedModel <- lm(data = PenguinOmit_reduced, flipper_length_mm ~
Delta.13.C..o.oo. + culmen_depth_mm + culmen_length_mm + body_mass_g)

vif(reducedModel)

summary(reducedModel)

anova(reducedModel, fullModel)
```

Residual Analysis

```
stdres(reducedModel)
studres(reducedModel)
```

```
rstudent(reducedModel)

range(stdres(reducedModel))

barplot(height = stdres(reducedModel), main = "Standardized Residuals", xlab =
"Index", ylab = "Standardized Resid", ylim=c(-6,6), cex.names = 0.8)
abline(h=3, col = "Red", lwd=2)
abline(h=-3, col = "Red", lwd=2)

range(studres(reducedModel))

barplot(height = studres(reducedModel), main = "Studentized Residuals", xlab =
"Index", ylab = "Studentized Resid", ylim=c(-6,6), cex.names = 0.8)
abline(h=3, col = "Red", lwd=2)
abline(h=-3, col = "Red", lwd=2)

range(rstudent(reducedModel))

barplot(height = rstudent(reducedModel),, main = "R Student Residuals", xlab = "Index",
ylab = "R Student Resid", ylim=c(-6,6), cex.names = 0.8)
abline(h=3, col = "Red", lwd=2)
abline(h=-3, col = "Red", lwd=2)
```

Influential Points Analysis

```
dfbetasPlots(reducedModel, intercept = TRUE)
influenceIndexPlot(reducedModel)
length(summary(influence.measures(reducedModel)))
```

QQ Plot and Histogram

```
par(mfrow=c(1,2))
hist(studres(reducedModel), breaks=10, freq=F, col="cornflowerblue",
    cex.axis=1.5, cex.lab=1.5, cex.main=2)
qqPlot(reducedModel)
```

Residuals v. Fitted Values

```
# residuals v. fitted values
```

```
residualPlot(reducedModel, type="rstudent", fitted = F, quadratic = F, col = 6, pch=16,
cex=1.5, cex.axis=1.5, cex.lab=1.5)
```

Box-Cox Transformation

```
bc <- boxcox(reducedModel, plotit = TRUE)
(bc.power <- bc$x[which.max(bc$y)])
```