ANALYSIS OF PALMER ARCHIPELAGO PENGUINS
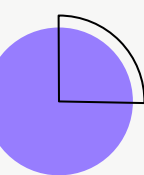
Team HAT: **H**oward Mach, **A**ndy Chandler, **T**hompson Pham

DATA SET INFORMATION

SIZE	344 observations with 17 variables, totaling 5,848 data points, detailing penguins in the Antarctic Peninsula
SOURCE	Samples from the Palmer Station, as a part of the Long Term Ecological Research Network in Antarctica working to analyze the penguins of 3 different islands on or near the Palmer Archipelago
RESPONSE VARIABLE	Flipper Length (mm)
PREDICTOR VARIABLES	Clutch Completion, Culmen Length (mm), Culmen Depth (mm), Body Mass (g), Delta 15 N (o/oo), Delta 13 C (o/oo), Sex, Species, Island, Date Egg
ANALYSIS GOAL	Our goal was predict the flipper length of each given penguin based on their varying characteristics

Data set viewing locations:
KAGGLE | GITHUB

Environmental Data Initiative:
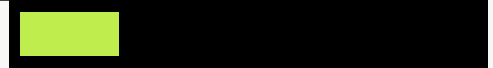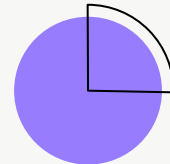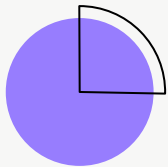**CHINSTRAP | GENTOO | ADÉLIE**
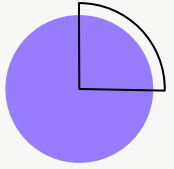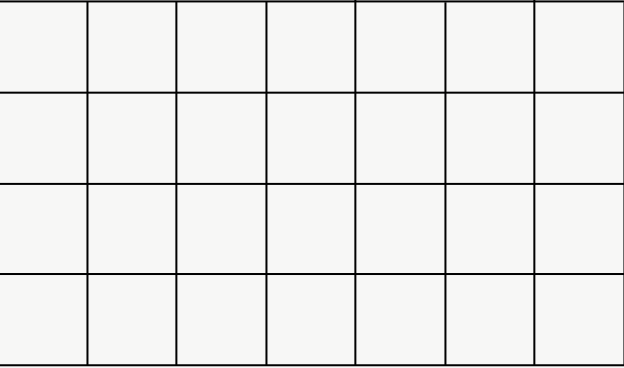
01

DATA SET CLEANING & EXPLORATORY ANALYSIS
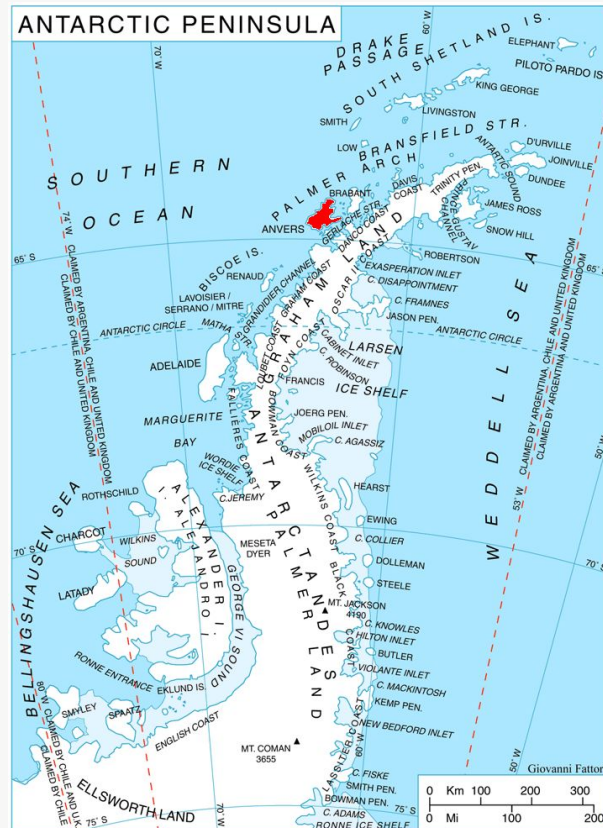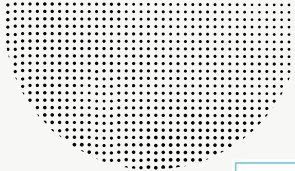
DATA CLEANING

- Removed Rows w/ NA
  - 19 Rows

- **17** Variables, of which we removed **5**
  - Study name
  - Sample number
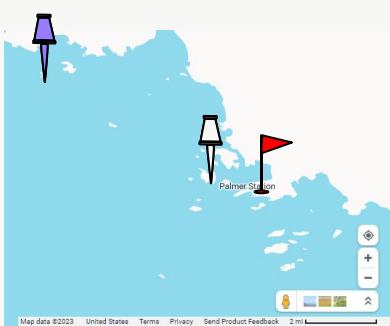 - Individual ID
  - Region
  - Comments

THE PALMER STATION

~125.32 miles North of the Antarctic Circle
64°46′27″S 64°03′10″W

ANTARCTIC PENINSULA

Dream Island

~6.86 miles from Palmer Station 64°44′S 64°14′W
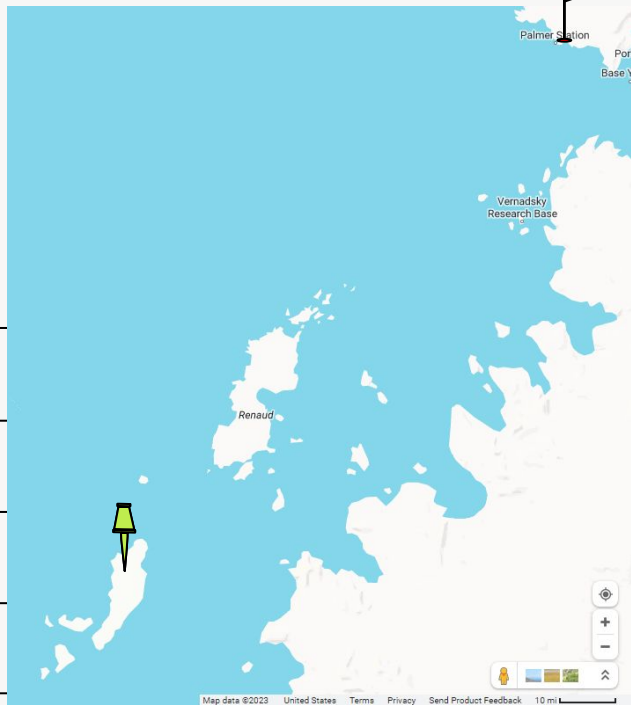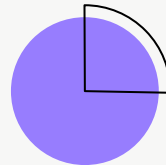
Torgersen Island

~1.27 miles from Palmer Station: 64°46′S 64°5′W

Biscoe Islands

~118.74 miles from Palmer Station 65°26′S 65°30′W

PENGUIN SPECIES

ChinStrap

Gentoo
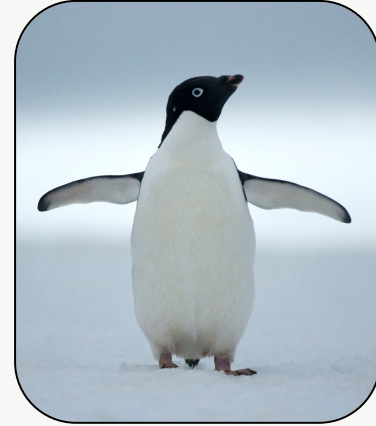
Adélie

CULMEN: RIDGE ALONG THE
TOP PART OF A BIRD'S BILL

CULMEN LENGTH

CULMEN
DEPTH
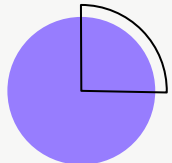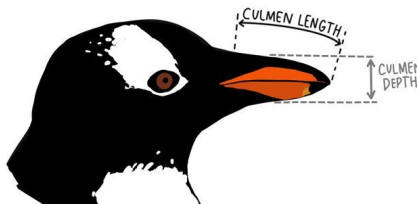
EXPLORATORY DATA ANALYSIS
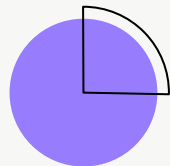
- Bimodal Distribution
- Two distinct groups

- Potential Correlation
- Adelie & Chinstrap

```
species      flipper length_mm
Adelie       172 - 210
Chinstrap    178 - 212
Gentoo       203 - 231
```

Distribution of Penguin Flipper Length

species

- Adelie
- Chinstrap
- Gentoo
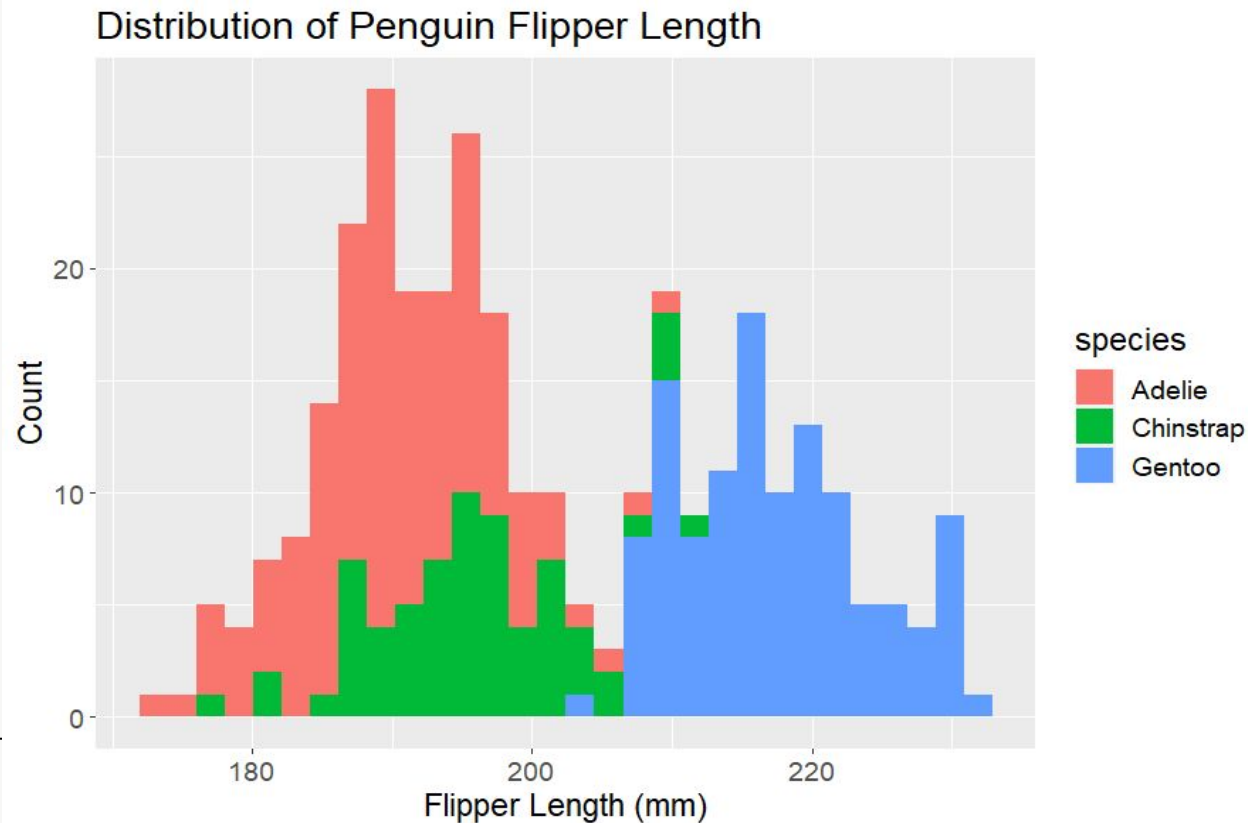
EXPLORATORY DATA ANALYSIS (CONTINUOUS)
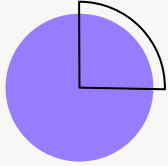
- Generally has a strong linear relationship
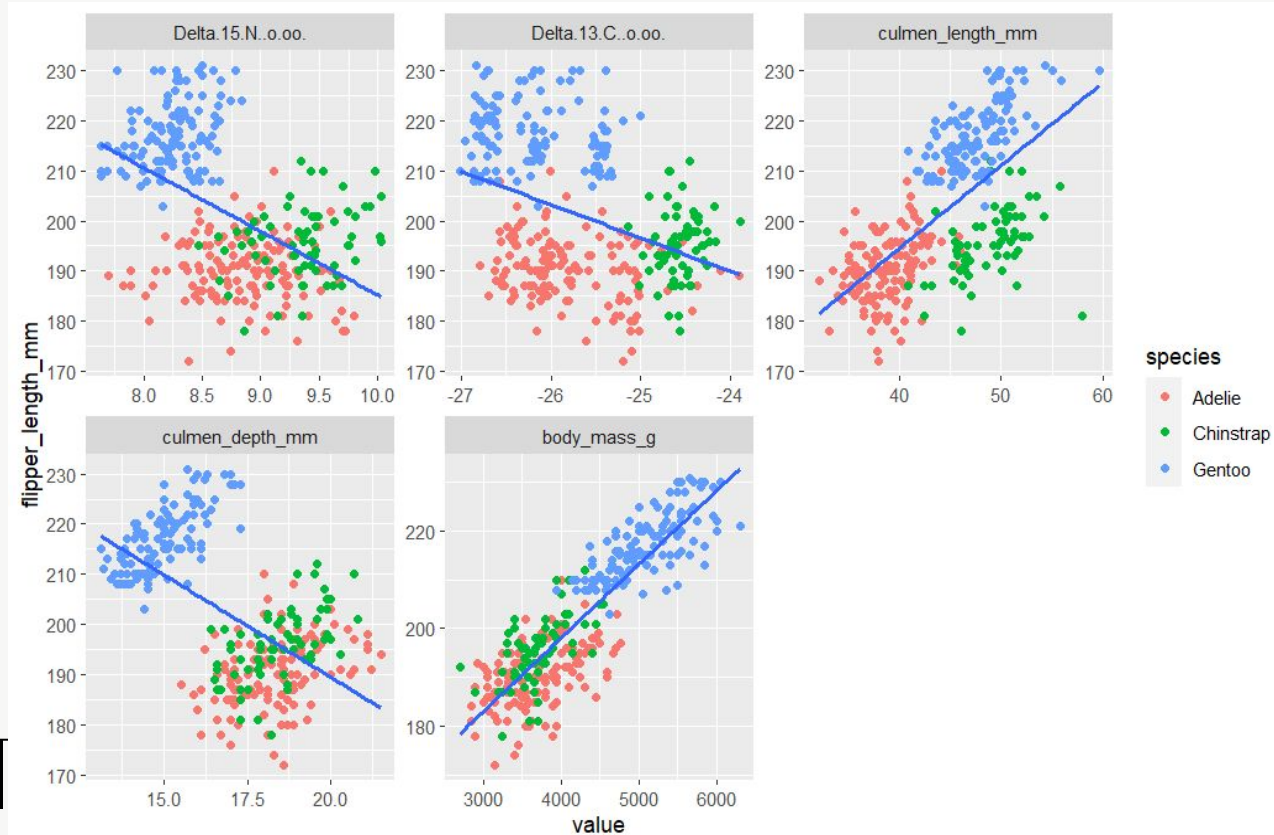
EXPLORATORY DATA ANALYSIS (DISCRETE)
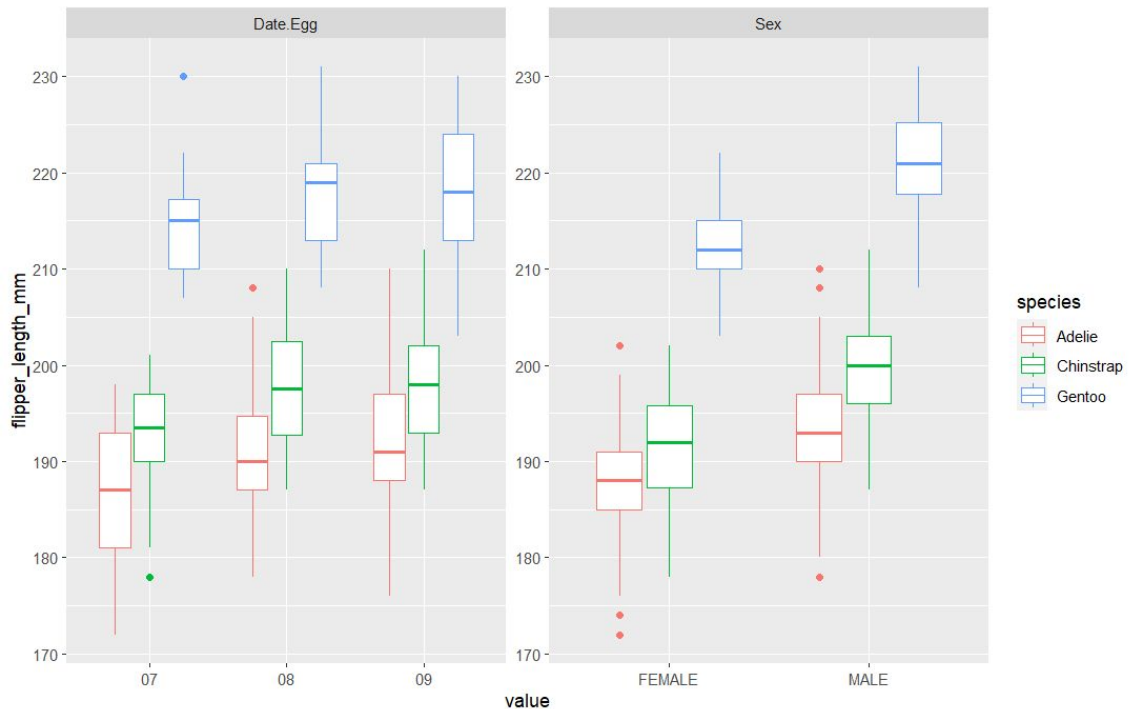
- Across discrete and continuous variables correlation between Adelie and Chinstrap

- Species has a significant factor

02

MODEL CONSTRUCTION & FITTING

MODEL CONSTRUCTION

```
> 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
```
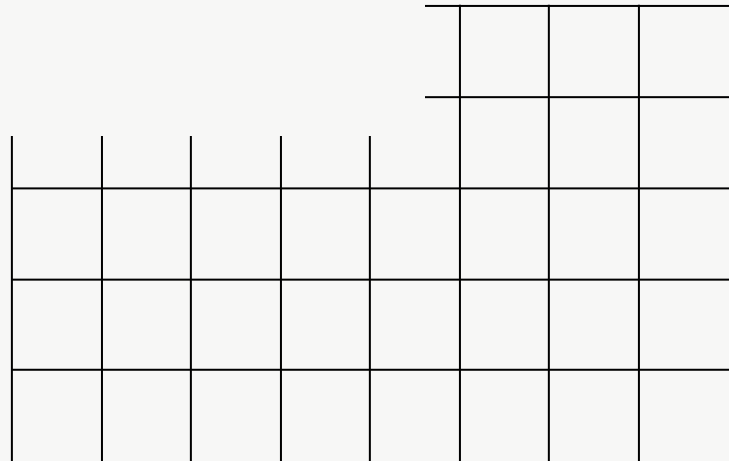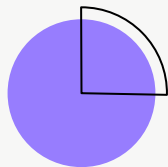
- **Delta 13, Culmen Depth, Culmen Length,** and **Body mass** are influential regressors, with the latter two having the highest influence

- **Delta 15** is the only variable without a strong influence to **Flipper Length**

VARIANCE INFLATION FACTORS

```
> vif(fullModel)
Delta.15.N..o.oo.  Delta.13.C..o.oo.    culmen_depth_mm   culmen_length_mm        body_mass_g
         2.256680           1.927347           1.719255           2.348709           2.977464
```

| Variable:	Delta 13	Culmen Depth	Culmen Length	Body Mass
VIF:	1.927347	1.719255	2.348709	2.977464

- No signs of multicollinearity
- No need to remove the remaining variables

```
> vif(reducedModel)
Delta.13.C..o.oo.    culmen_depth_mm   culmen_length_mm        body_mass_g
         1.819519           1.425882           2.259734           2.603961
```

Variable:	Delta 13	Culmen Depth	Culmen Length	Body Mass
VIF:	1.819519	1.425882	2.259734	2.603961
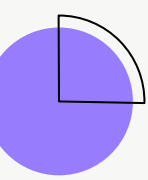
ANOVA COMPARISON

```
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
```

- After removing **Delta 15** and an outlier point we had we created our reduced model
- Adj R^2 = 0.8527

```
> 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
```
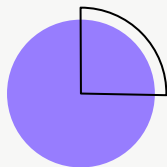
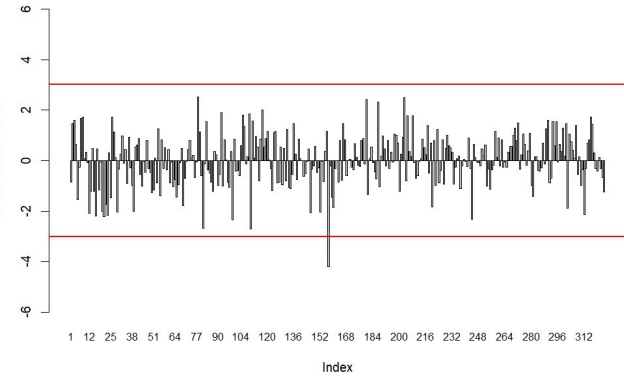With a P value significantly greater than α = 0.5, our reduced model is better than the full model
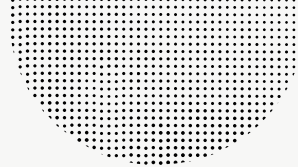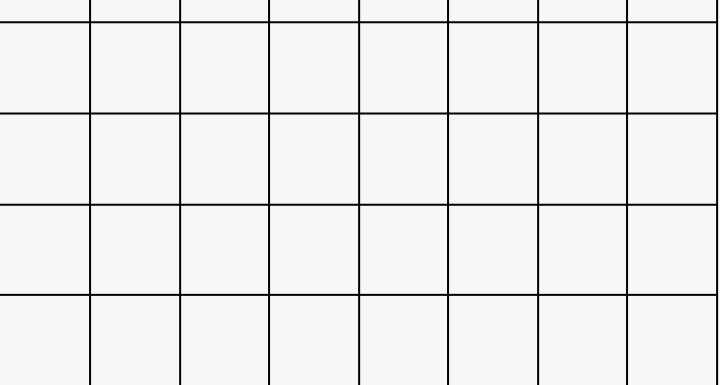
TRANSFORMATION, RESIDUAL ANALYSIS & **INFLUENTIAL** POINTS
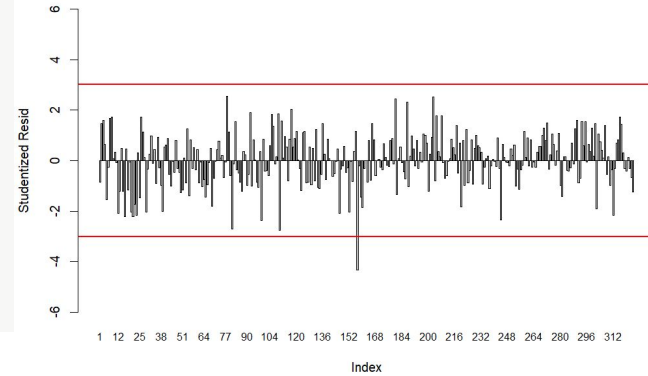
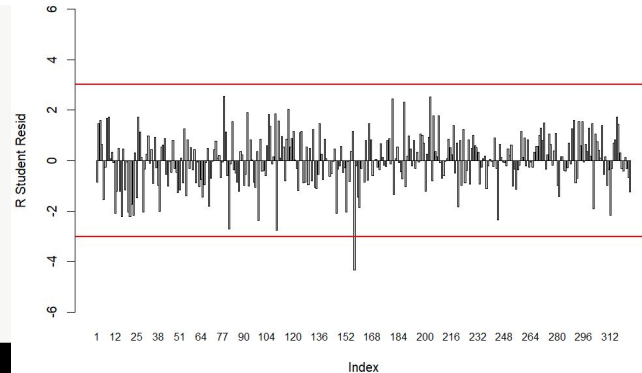# RESIDUAL ANALYSIS - Full Model

Standardized Residuals

Studentized Residuals

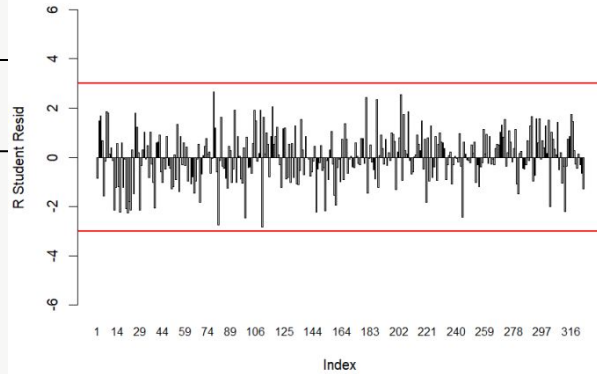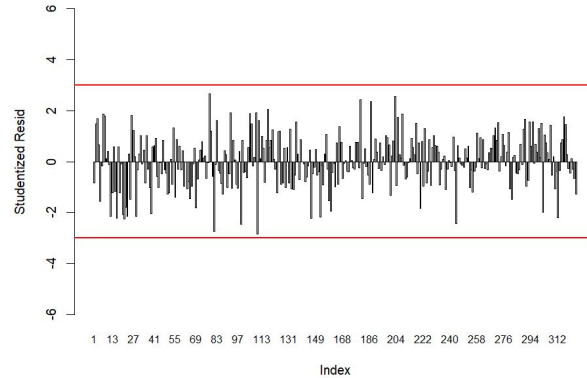
R Student Residuals

RESIDUAL ANALYSIS - Reduced Model
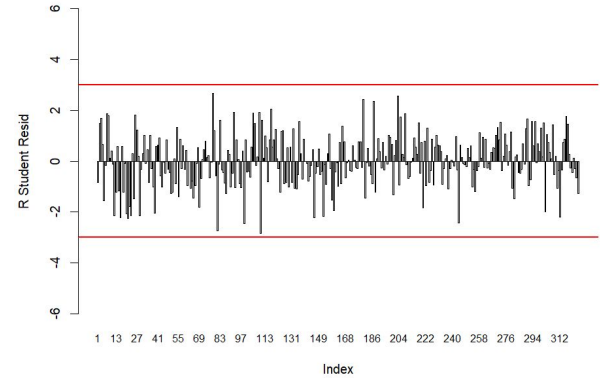
INFLUENTIAL POINTS ANALYSIS

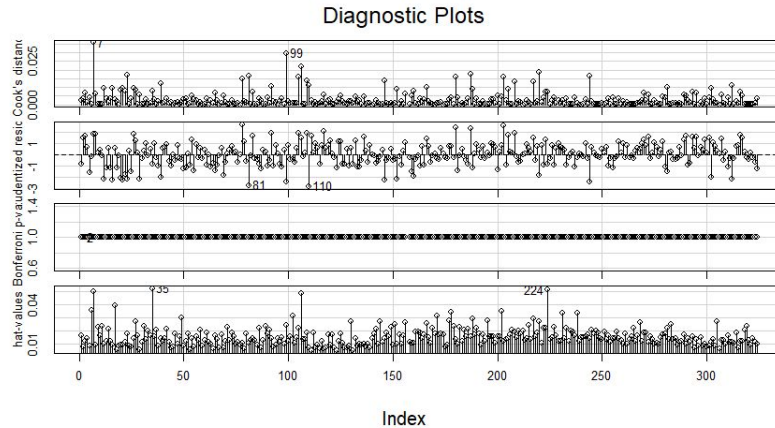Diagnostic Plots

dfbetas Plots
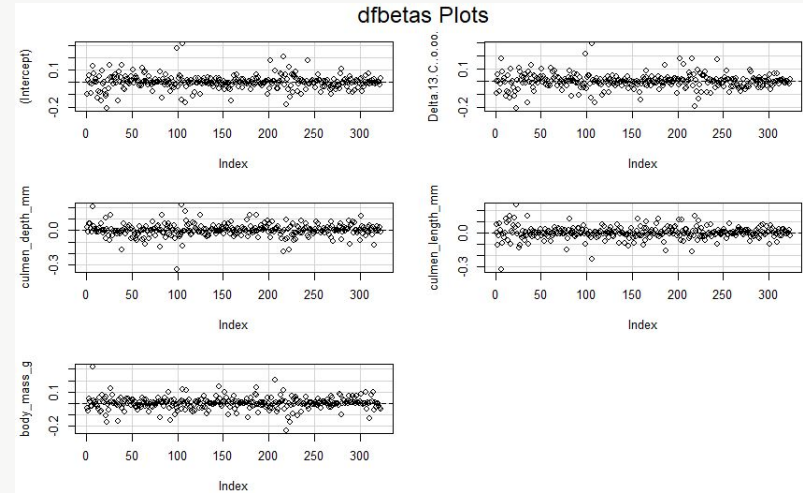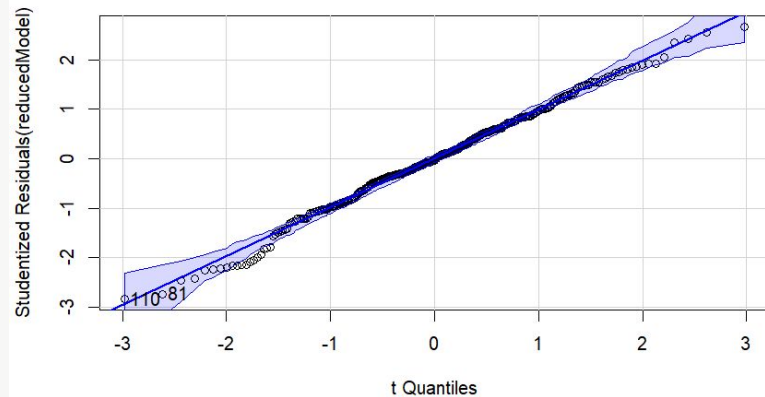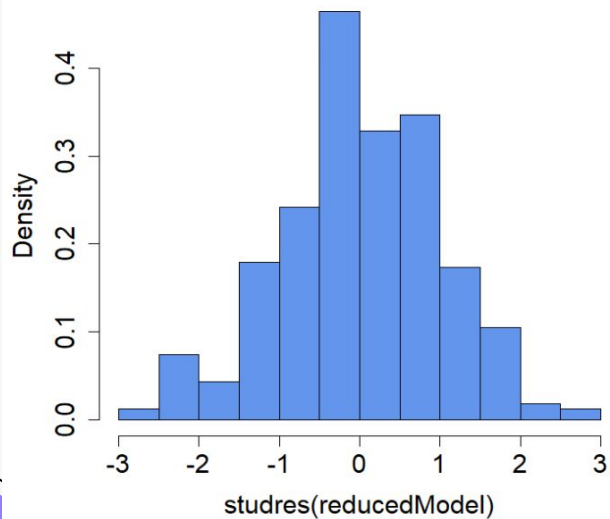
19 influential points, which makes up 5.88% of the dataset

Hat values: 4
DFBETAS: 9
DFFITS: 2
COVRATIO: 17

Influential points:
- Observation #7
- Observation #99
- Observation #106

TRANSFORMATION

TRANSFORMATION

The fitted values are spread out, but converge around the zero line.

Residuals v. Fitted Values
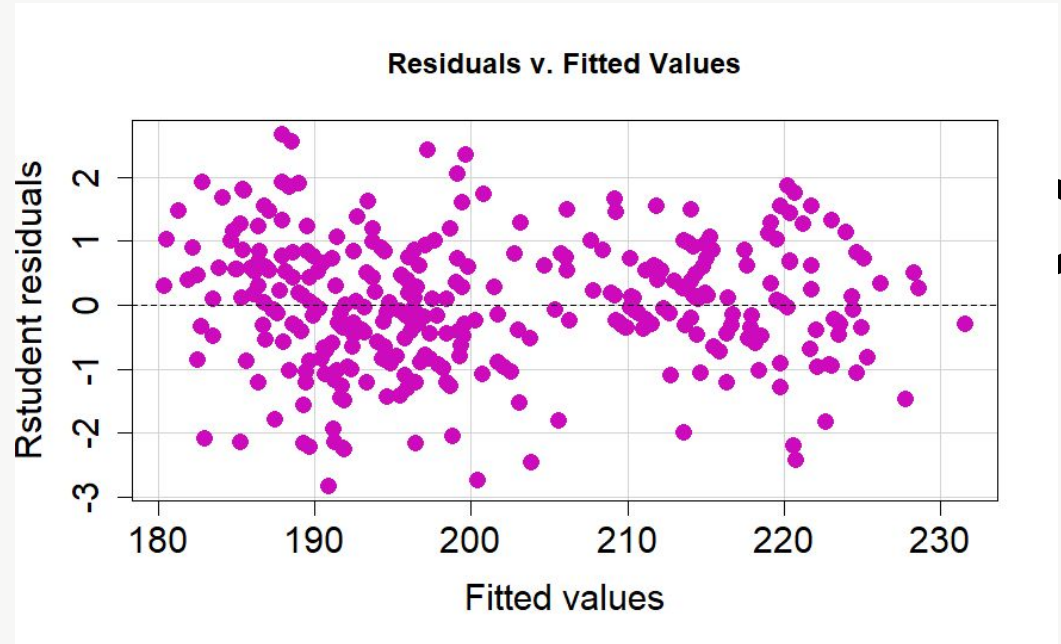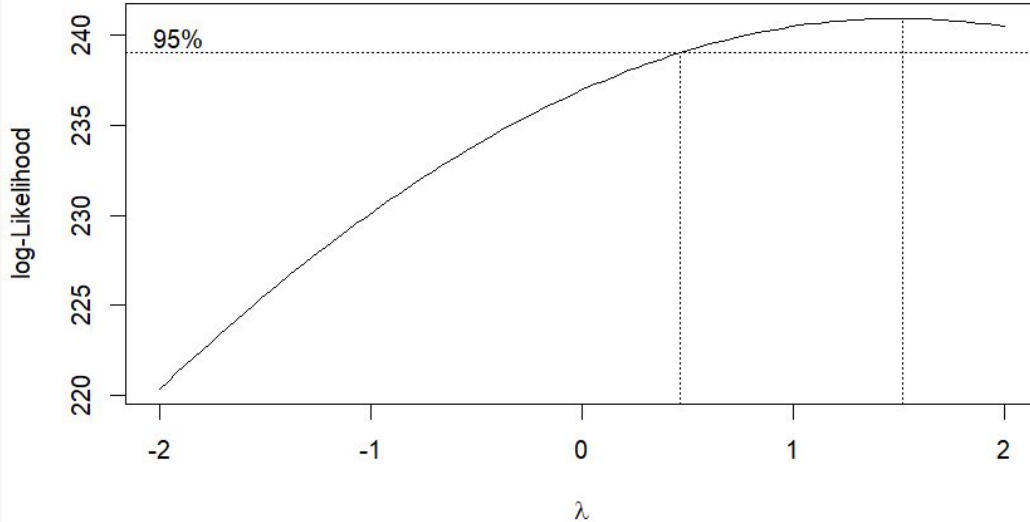
TRANSFORMATION

λ = 1.515152

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

Y = 74.9158237 - 2.8069258 D13_C - 1.2942930 CulD_mm + 0.8636770 CulL_mm + 0.0091095 BodyM_g

Future Goals:
- In the future we hope to possibly 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
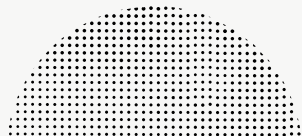
Citations

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 (Accessed 2023-12-04).

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 (Accessed 2023-12-04).

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 (Accessed 2023-12-04).
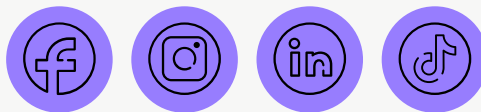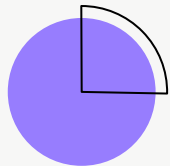
Gorman KB, Williams TD, Fraser WR (2014) Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins (Genus Pygoscelis). PLoS ONE 9(3): e90081. doi:10.1371/journal.pone.0090081. https://www.kaggle.com/datasets/parulpandey/palmer-archipelago-antarctica-penguin-data/data

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

THANKS

Do you have any questions?