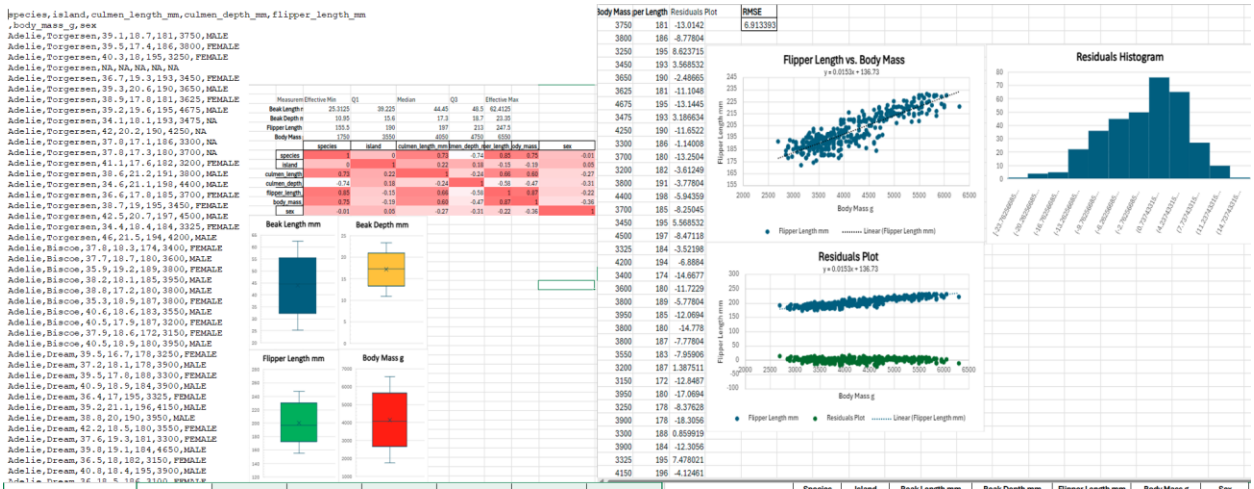
Tim Rishell-EXCEL PROJECT

An Analysis of Penguin Growth

By: Tim Rishell

Project Date: 6/23/2024



	species	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex
COUNT, cells with numbers	0	0	342	342	342	342	0
COUNTA, cells not empty	344	344	344	344	344	344	344
COUNTIF, cells with "NA"	0	0	2	2	2	2	10
AVERAGE, ignores "NA"	N/A	N/A	43.92	17.15	200.92	4201.75	N/A
STDEV, ignores "NA"	N/A	N/A	5.46	1.97	14.06	801.95	N/A
VAR	N/A	N/A	29.81	3.90	197.73	643131.08	N/A
MIN	N/A	N/A	32.1	13.1	172	2700	N/A
MAX	N/A	N/A	59.6	21.5	231	6300	N/A
Row 2 as example	Adelie	Torgersen	39.1	18.7	181	3750	MALE
Data type	Qualitative	Qualitative	Quantitative	Quantitative	Quantitative	Quantitative	Qualitative
Numerical or Categorical	Categorical	Categorical	Numerical	Numerical	Numerical	Numerical	Categorical
Continuous or Discrete	N/A	N/A	Continuous	Continuous	Continuous	Continuous	N/A

	Species	Island	Beak Length mm	Beak Depth mm	Flipper Length mm	Body Mass g	Sex
COUNT, cells with numbers	342	342	342	342	342	342	342
COUNTA, cells not empty	342	342	342	342	342	342	342
COUNTIF, cells with "NA"	0	0	0	0	0	0	0
AVERAGE, ignores "NA"	N/A	N/A	43.92	17.15	200.92	4201.75	N/A
STDEV, ignores "NA"	N/A	N/A	5.46	1.97	14.06	801.95	N/A
VAR	N/A	N/A	29.81	3.90	197.73	643131.08	N/A
MIN	1	1	32.1	13.1	172	2700	0
MAX	3	3	59.6	21.5	231	6300	2
Row 2 as example	1	1	39.1	18.7	181	3750	1
Data type	Qualitative	Qualitative	Quantitative	Quantitative	Quantitative	Quantitative	Qualitative
Numerical or Categorical	Categorical	Categorical	Numerical	Numerical	Numerical	Numerical	Categorical
Continuous or Discrete	N/A	N/A	Continuous	Continuous	Continuous	Continuous	N/A
MODE, numerical fields only	1	2	41.1	17	190	3800	1

Table of Contents

EXECUTIVE SUMMARY 3
DATA SET 3
DATA EXPLORATION 4
DATA PREPARATION 5
PRELIMINARY ANALYSIS 6
ANALYSIS 8
CONCLUSIONS 10

EXECUTIVE SUMMARY

The following project is a showcase of my acquired skills in Excel and my knowledge of basic statistics. For this project, Excel was used to convert .csv data about the measurements of penguins into a spreadsheet. I then performed multiple steps of data cleaning and preparation prior to using various forms of charts to test and create a linear regression model. This linear regression model was used to see the positive relationship between body mass and flipper length growth. After further analysis of the data, linear regression models were created to reflect the individual species of penguins measured and the aforementioned positive relationship. Final conclusions showed that the models were all accurate for predictions, however, the specific species breakdowns were more accurate than the generalized model. Additionally, the data revealed that while penguins generally grow in the same way, they have species specific growth rates, along with other details that can affect growth that could be studied.

DATA SET

Here you will find a description and analysis of the dataset named "penguins_size.csv". This dataset will show measurements taken of 3 species of penguins from 3 island locations in Antarctica. The data recorded is as follows:

- The species of the penguin.
- The island the penguin was found and measured at.
- The length of the beak in millimeters of the penguins from the base of the beak to the tip.
- The vertical height of the beak in millimeters of the penguins at the base of the beak.
- The length of the flippers in millimeters from the shoulder joint to the tip of the flipper.
- The body mass in grams of the penguin.
- The sex of the penguin.

These measurements can provide insight to study and predict details of the penguins such as their size, age, sex, feeding habits, health, species or subspecies, physical development, breeding status, and more. This data can be used to further understand penguins and to strengthen conservation efforts.

DATA EXPLORATION

	species	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex
COUNT, cells with numbers	0	0	342	342	342	342	0
COUNTA, cells not empty	344	344	344	344	344	344	344
COUNTIF, cells with "NA"	0	0	2	2	2	2	10
AVERAGE, ignores "NA"	N/A	N/A	43.92	17.15	200.92	4201.75	N/A
STDEV, ignores "NA"	N/A	N/A	5.46	1.97	14.06	801.95	N/A
VAR	N/A	N/A	29.81	3.90	197.73	643131.08	N/A
MIN	N/A	N/A	32.1	13.1	172	2700	N/A
MAX	N/A	N/A	59.6	21.5	231	6300	N/A
Row 2 as example	Adelie	Torgersen	39.1	18.7	181	3750	MALE
Data type	Qualitative	Qualitative	Quantitative	Quantitative	Quantitative	Quantitative	Qualitative
Numerical or Categorical	Categorical	Categorical	Numerical	Numerical	Numerical	Numerical	Categorical
Continuous or Discrete	N/A	N/A	Continuous	Continuous	Continuous	Continuous	N/A
MODE, numerical fields only	N/A	N/A	41.1	17	190	3800	N/A

Visual inspection of the data shown in the table will reveal that there is a positive linear relationship between the measurements taken of the penguins. Initial impressions of the data shown would also suggest a normal distribution of the data, further analysis will reveal if there is a skew.

Tim Rishell-EXCEL PROJECT

DATA PREPARATION

Below you will find a table reflecting a cleaned version of the dataset. The following key describes the values that were converted for analysis purposes.

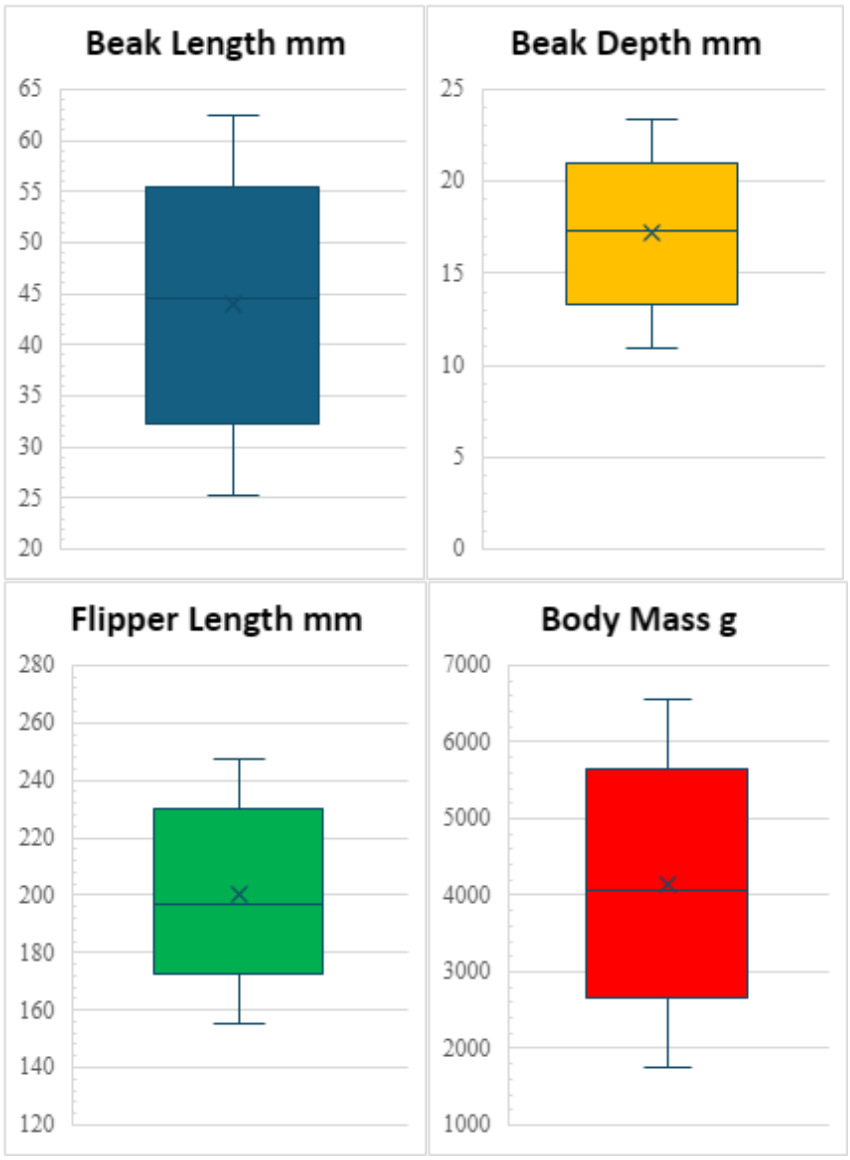
Key:	Species	Island	Sex
	Adelie = 1 Chinstrap = 2 Gentoo = 3	Torgersen = 1 Biscoe = 2 Dream = 3	Male = 1 Female = 2 UNKNOWN = 0

	Species	Island	Beak Length mm	Beak Depth mm	Flipper Length mm	Body Mass g	Sex
COUNT, cells with numbers	342	342	342	342	342	342	342
COUNTA, cells not empty	342	342	342	342	342	342	342
COUNTIF, cells with "NA"	0	0	0	0	0	0	0
AVERAGE, ignores "NA"	N/A	N/A	43.92	17.15	200.92	4201.75	N/A
STDEV, ignores "NA"	N/A	N/A	5.46	1.97	14.06	801.95	N/A
VAR	N/A	N/A	29.81	3.90	197.73	643131.08	N/A
MIN	1	1	32.1	13.1	172	2700	0
MAX	3	3	59.6	21.5	231	6300	2
Row 2 as example	1	1	39.1	18.7	181	3750	1
Data type	Qualitative	Qualitative	Quantitative	Quantitative	Quantitative	Quantitative	Qualitative
Numerical or Categorical	Categorical	Categorical	Numerical	Numerical	Numerical	Numerical	Categorical
Continuous or Discrete	N/A	N/A	Continuous	Continuous	Continuous	Continuous	N/A
MODE, numerical fields only	1	2	41.1	17	190	3800	1

The cleaned version of the dataset and subsequent table reveals:

- The Adelie species is the most frequently measured penguin species in this set.
- There were more penguins measured at Biscoe Island than the other islands.
- The most commonly occurring sex amongst the penguins measured is Male.

PRELIMINARY ANALYSIS



The boxplots above reveal the data that will be used for modeling and later analysis.

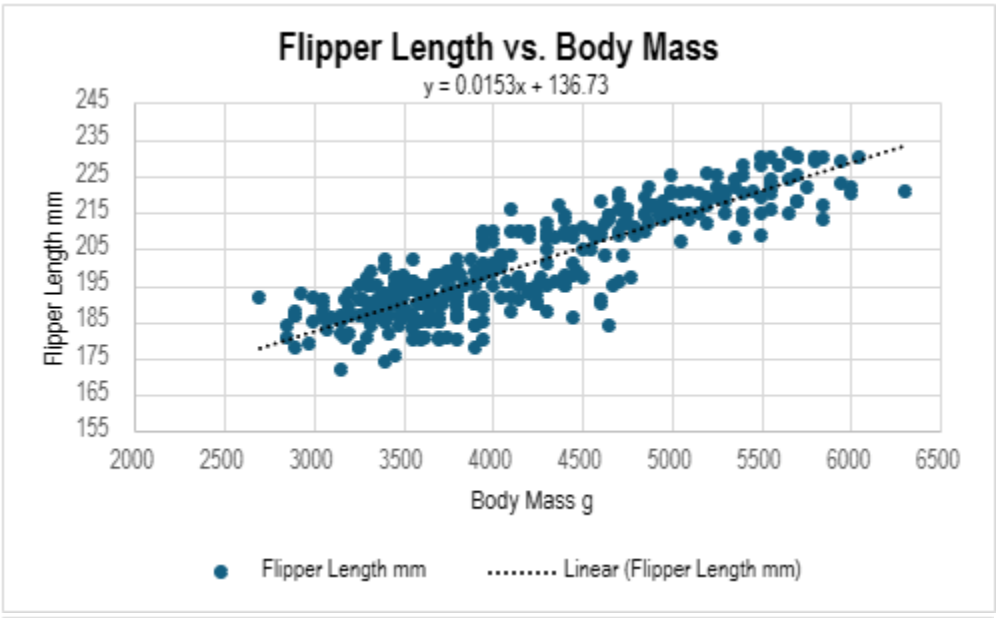
Penguins are natural swimmers and have evolved to use this ability to their advantage. They hunt fish with their beaks, which contain many rows of needle-like spines called papillae. The measurements in this dataset align with the features that make penguins so fierce in the water. As penguins age and grow, their bodies adapt to their hunting grounds by increasing the length of their flippers, enhancing their speed in water. The size and shape of the penguin's beak, their primary hunting tool, are strongly tied to flipper length and body mass. Generally, the heavier the penguin and the longer the flippers, the longer the beak. Most notably, as flipper length, body mass, and beak length increase, the depth of the beak decreases. This narrow beak design allows penguins to be more hydrodynamic, enabling them to catch fast-moving fish efficiently.

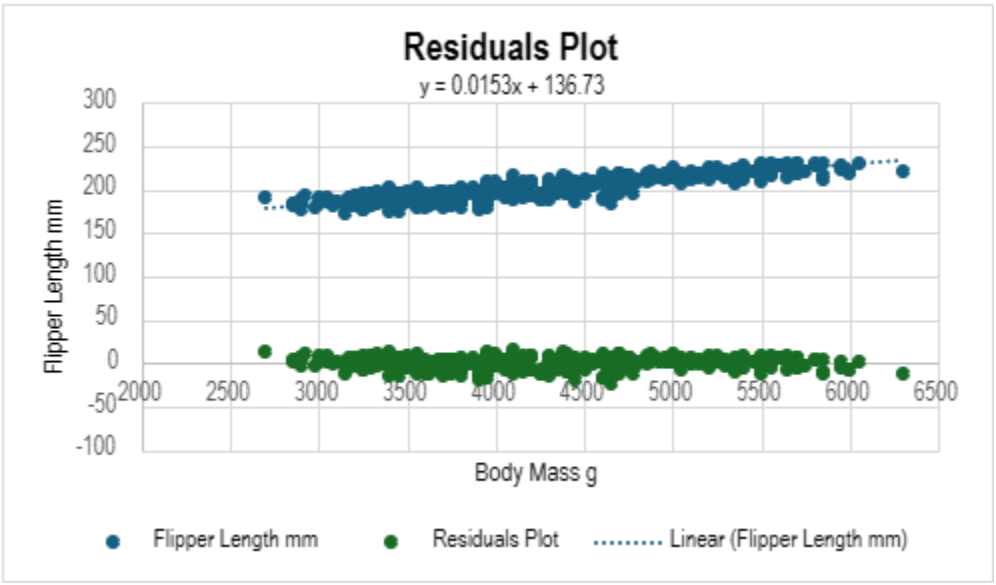
The following correlation chart supports this analysis.

	species	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex
species	1	0	0.73	-0.74	0.85	0.75	-0.01
island	0	1	0.22	0.18	-0.15	-0.19	0.05
culmen_length_mm	0.73	0.22	1	-0.24	0.66	0.60	-0.27
culmen_depth_mm	-0.74	0.18	-0.24	1	-0.58	-0.47	-0.31
flipper_length_mm	0.85	-0.15	0.66	-0.58	1	0.87	-0.22
body_mass_g	0.75	-0.19	0.60	-0.47	0.87	1	-0.36
sex	-0.01	0.05	-0.27	-0.31	-0.22	-0.36	1

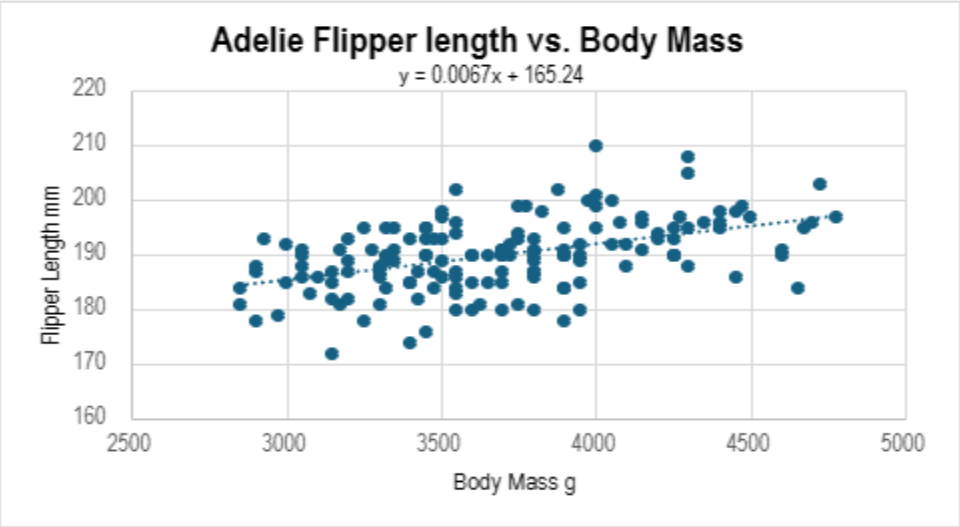
ANALYSIS

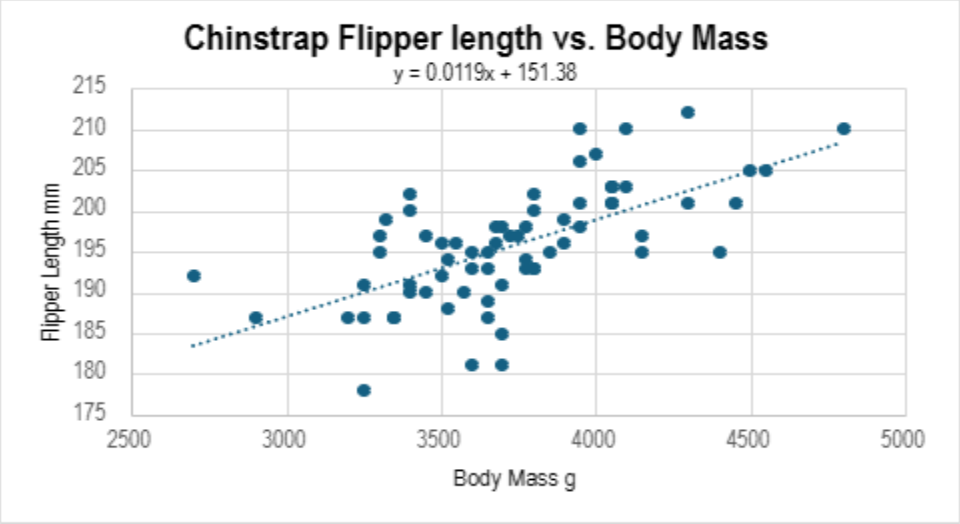
The main aspect of growth in penguins is the body mass and flipper length, with these variables then being correlated to other areas of growth. As shown in the following linear regression models, the correlation between body mass and flipper length is strong. The fitness of the linear regression model can be verified by the accompanying residuals plot showing the near 0 mean and normality of the residuals.

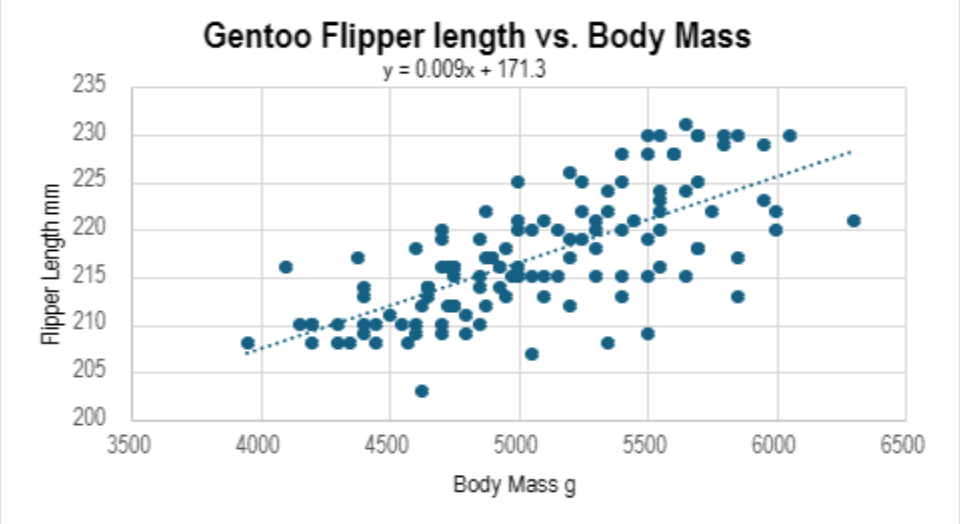




The accuracy of the model changes when looking at specific species of penguins, shown here.







CONCLUSIONS

The following calculations explain the difference in the accuracy of the linear regression models. The RMSE calculated for the models reveals how accurate the model is to be used for predictions about the respective population. The lower the RMSE value, the more accurate the predictions about the population will be when using that model.

General RMSE	Adelie RMSE	Chinstrap RMSE	Gentoo RMSE
6.91	5.80	5.51	4.63

The RMSE values here are low enough to determine that the model is accurate enough to be used for general predictions about the population. However, the accuracy of the model gets stronger when focusing on specific species of penguins. This shows that while penguins develop in a similar way, their overall development varies based on species, and likely other factors that could be further investigated.