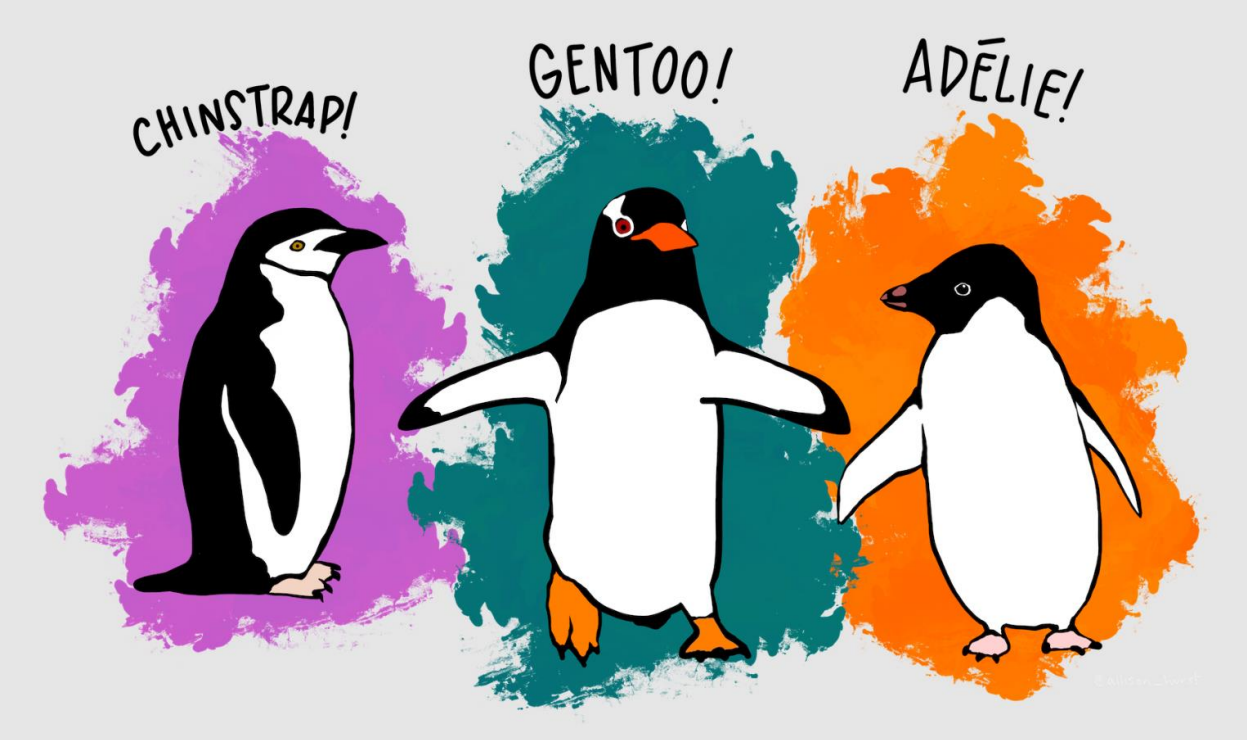
Palmer Penguins Review: a Predictive Model Approach.

Alec Brooks

8/1/2024



Executive Summary

The Palmer Penguin dataset is widely used in academic settings due to its variety and scope, making it a good tool for exploring different types of statistical analysis. The dataset includes seven different metrics that can be analyzed for various insights. This report focuses on the correlation between penguin Body Mass and Flipper Length. This analysis shows that Flipper Length can be used as a reliable predictor of penguin Body Mass, using the formula $y = 50.153x - 5872.1$ within acceptable levels of accuracy. However, these findings are preliminary, and further analysis could reveal an additional wealth of other connections within the dataset.

Data Set

In this report, we will be examining the "palmer-penguins" dataset, provided by Allison Horst, a data scientist and artist on GitHub. The data was collected from 07 - 09 by Dr. Kristen Gorman, while working with the US Long Term Ecological Research Network. This dataset includes three categorical variables. species, island, and sex, along with four numerical variables. culmen_length_mm, culmen_depth_mm, flipper_length_mm, and body_mass_g. Together, these variables form the palmer-penguins dataset, which is largely used in training and classroom settings due to its simplicity and variety of variable types.

Data Exploration

Upon initial exploration of our dataset, we observe that all present numerical variables appear to be consistent with relatively low deviation.

Transposition Table	MIN	MAX	AVERAGE	STDEV	VAR
culmen_length_mm	32.1	59.6	43.9	5.5	29.8
culmen_depth_mm	13.1	21.5	17.2	2.0	3.9
flipper_length_mm	172.0	231.0	200.9	14.1	197.7
body_mass_g	2700.0	6300.0	4201.8	802.0	643131.1

As always, this exploration includes examining missing values. By doing this, we uncover three important facts that may be relevant to our analysis:

1. The dataset includes 344 total rows.
2. The "sex" variable has one missing value and 10 NAs.
3. Two records in our dataset include species and island values but are missing all numerical data.

	sex
COUNT cells with #s	0
COUNTA (nonempty)	343
COUNTIF "NA"	10

	species	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex
COUNTIF "NA"	0	0	2	2	2	2	10

These inconsistencies may become relevant in our analysis and will need to be addressed.

Data Preparation

To begin with data preparation, we removed rows containing NAs and numerically encoded the categorical variables, which grants more flexibility in our analysis.

	species	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex
COUNT cells with #s	0	0	342	342	342	342	0
COUNTA (nonempty)	344	344	344	344	344	344	343
COUNTIF "NA"	0	0	2	2	2	2	10
AVERAGE			43.9	17.2	200.9	4201.8	
STDEV			5.5	2.0	14.1	802.0	
VAR			29.8	3.9	197.7	643131.1	
MIN	0	0	32.1	13.1	172.0	2700.0	0
MAX	0	0	59.6	21.5	231.0	6300.0	0
Row 2	Adelie	Torgersen	39.1	18.7	181	3750	MALE
Data type	Text	Text	Number	Number	Number	Number	Text
Numerical or Categorical	Categorical	Categorical	Numerical	Numerical	Numerical	Numerical	Categorical
Continuous or Discrete	Discrete	Discrete	Continuous	Continuous	Continuous	Continuous	Discrete
MODE			41.1	17	190	3800	

One benefit of this encoding is that it makes it much easier to build a correlation matrix.

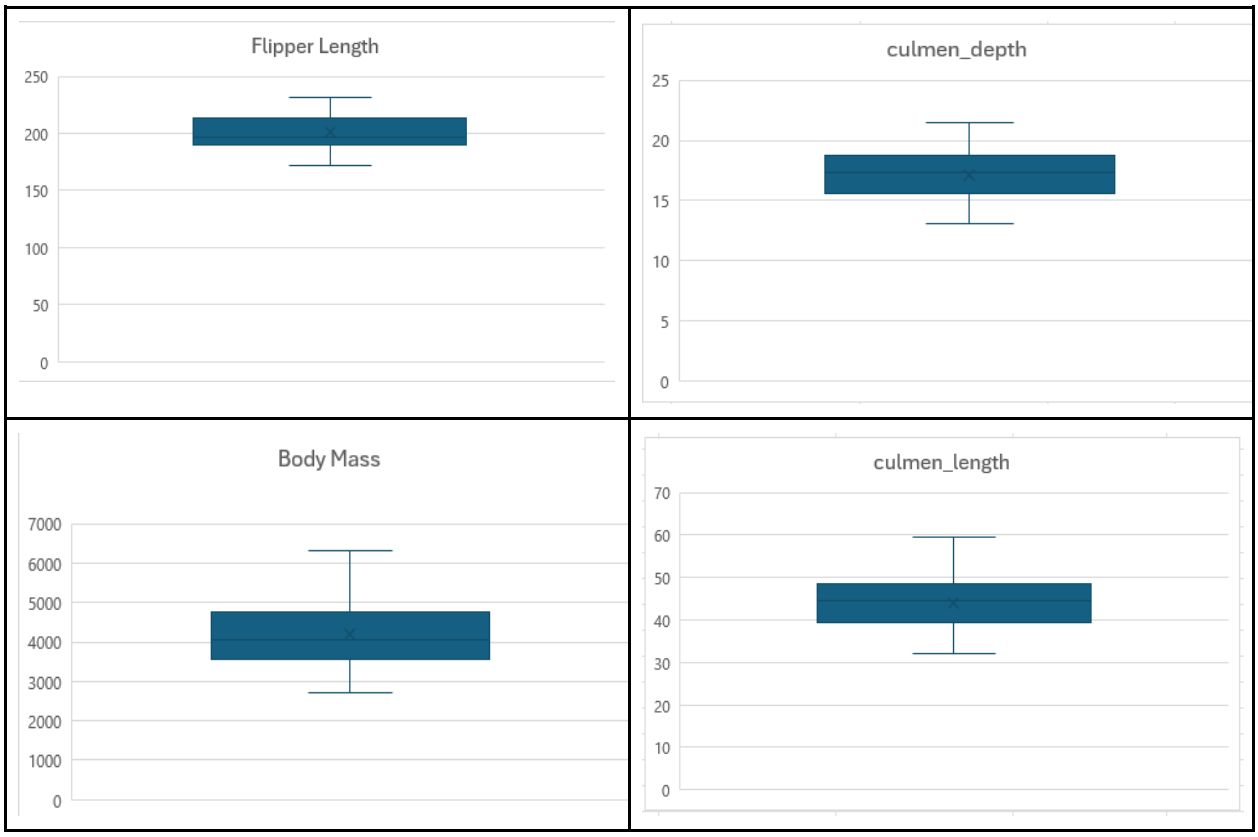
Correlation Matrix							
	species	island	culmen_length	culmen_depth	flipper_length	body_mass	sex
species	1.00	-0.01	0.73	-0.74	0.85	0.75	-0.01
island	-0.01	1.00	0.21	0.19	-0.16	-0.20	-0.01
culmen_length	0.73	0.21	1.00	-0.23	0.65	0.59	-0.34
culmen_depth	-0.74	0.19	-0.23	1.00	-0.58	-0.47	-0.37
flipper_length	0.85	-0.16	0.65	-0.58	1.00	0.87	-0.26
body_mass	0.75	-0.20	0.59	-0.47	0.87	1.00	-0.42
sex	-0.01	-0.01	-0.34	-0.37	-0.26	-0.42	1.00
Highest Correlation:	flipper_length : 0.851	culmen_length : 0.212	species : 0.731	island : 0.19	body_mass : 0.873	flipper_length : 0.873	island : -0.006
Lowest Correlation:	culmen_depth : -0.74	body_mass : -0.202	sex : -0.344	species : -0.74	culmen_depth : -0.578	culmen_depth : -0.472	body_mass : -0.425

Preliminary Analysis

	species_New	island_New	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex_New
COUNT cells with #s	333	333	333	333	333	333	333
COUNTA (nonempty)	333	333	333	333	333	333	333
COUNTIF "NA"	0	0	0	0	0	0	0
AVERAGE	1.9	2.2	44.0	17.2	201.0	4207.1	1.5
STDEV	0.9	0.7	5.5	2.0	14.0	805.2	0.5
VAR	0.8	0.5	29.9	3.9	196.4	648372.5	0.3
MIN	1	1	32.1	13.1	172	2700	1
MAX	3	3	59.6	21.5	231	6300	2
MODE	1	2	41.1	17	190	3800	1

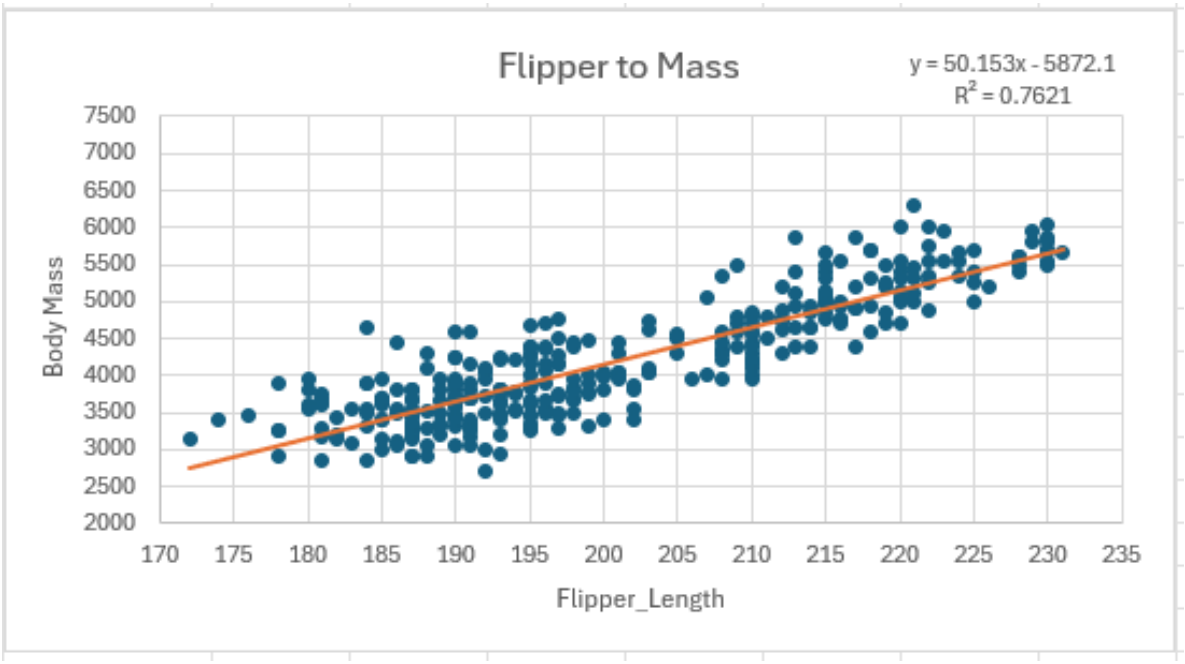
	culmen_length	culmen_depth	flipper_length	body_mass
Min	32.1	13.1	172	2700
Max	59.6	21.5	231	6300
Q1	39.5	15.6	190	3550
Q3	48.6	18.7	213	4775
IQR	9.1	3.1	23	1225
Effective Min	25.85	10.95	155.5	1712.5
Effective Max	62.25	23.35	247.5	6612.5

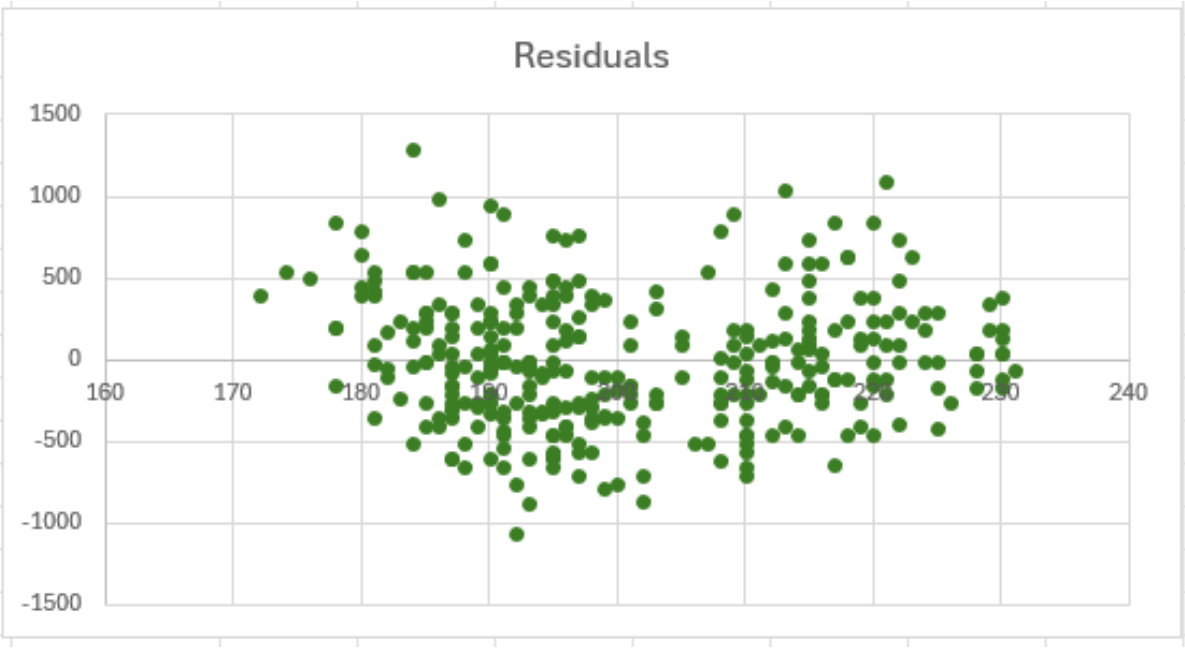
If we look at each quartile of our measurement variables it reveals a significant range and variability, with body mass showing the highest variation. Flipper length and body mass display moderate variability, making them suitable predictors for a linear regression model.



Analysis

If we plot Flipper Length as our independent variable and Body Mass as our dependent variable, we can see a consistent pattern. The R^2 score of 0.76 indicates that approximately 76% of the variance in Body Mass can be explained by Flipper Length. After we plot the residuals, we find a relatively low Mean Residual of 0.06.





Conclusion

Further, The Standard Error is 393.34. Considering that the range of the dependent variable (Body Mass) is 6300, this Standard Error is relatively low, representing about 6.24% of the Body Mass range. This indicates that the model is reasonably accurate, making Flipper Length an acceptable predictor of Body Mass. This project serves as a good case study for the importance of using methods such as plotting and analyzing residuals to verify findings.