Representative Survey of Antarctic Penguins

Nicholas Gunter

December 6, 2025

1 Data Cleaning

1.1 Loading and Removing Missing Data

To load the data, use csv reader and split into lists for each parameter. If any of the numerical values are missing (culmen length, culmen depth, body mass, or flipper length), skip that entry. This is an aggressive cleaning strategy as it deletes more information than it needs to. For example, a penguin missing a body mass measurement can still contribute to data concerning flipper length. The advantage is that it is a simple heuristic to implement and it keeps the length of all the lists consistent for easy analysis. The size of the data set is an important consideration for this. Since the data set is rather large and there are many penguins with complete measurements this solution can be implemented.

1.2 Drop Outliers

Similarly to deleting missing data, the outliers were dropped from the data. The goal of this study is to derive representative values for the penguin populations in question. Removing the outliers makes the distribution artificially narrow, but improves the accuracy of the mean as it will not be influenced by large outliers. For future analysis that is more concerned with the spread of penguin measurements, outliers are more important to keep.

1.3 Standardizing Categorical Variables

Similarly to removing missing numerical data, penguins with missing categorical variables were removed. Only a few penguins were missing sex information, while the species and island names were all recorded. The case and grammar of the categorical variables were already consistent in the original CSV, so nothing needed to be done to ensure that they were comparable by code. After all of the cleaning, there were 310 penguins left in the cleaned CSV.

2 Exploratory Data Analysis

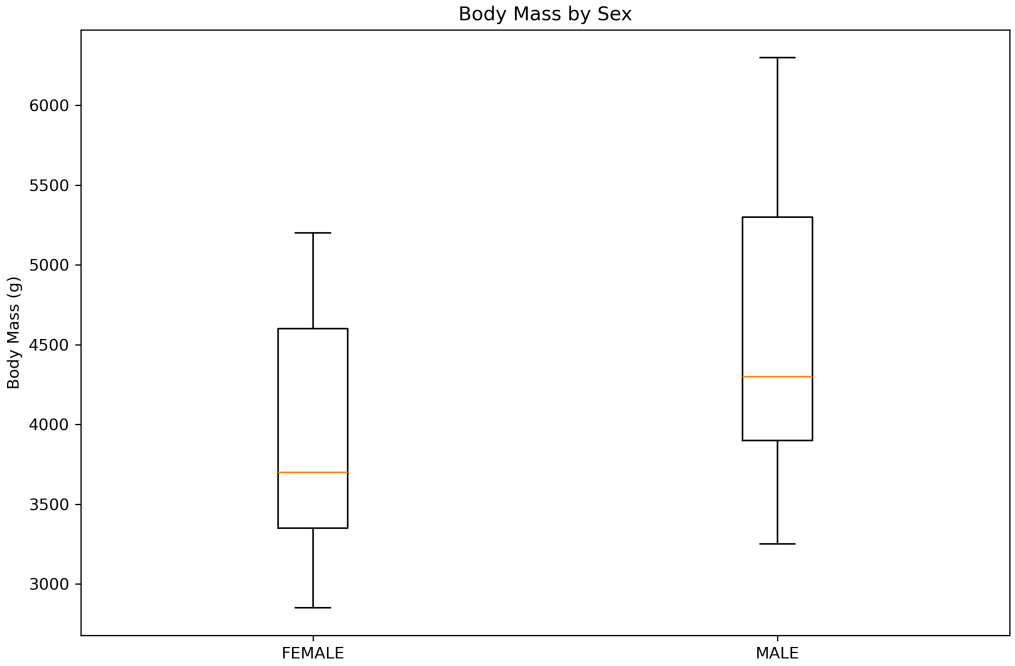

Figure 1: Box plot of body mass by sex

This figure shows that the median body mass of male penguins is higher than the mean body mass of female penguins. Both box plots are skewed upwards which indicates that most of the penguins with lower body mass than the median are close to the median.

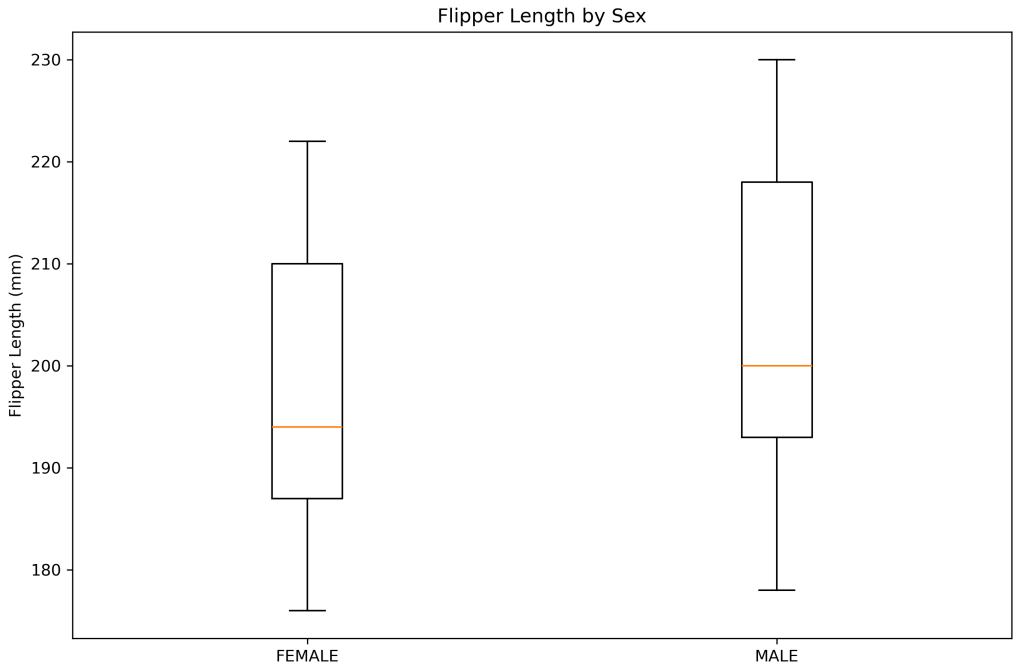

Figure 2: Box plot of flipper length by sex

The box plot of flipper length compared to sex shows similar trends to body mass. This likely indicates that flipper length and body mass are correlated in some way.

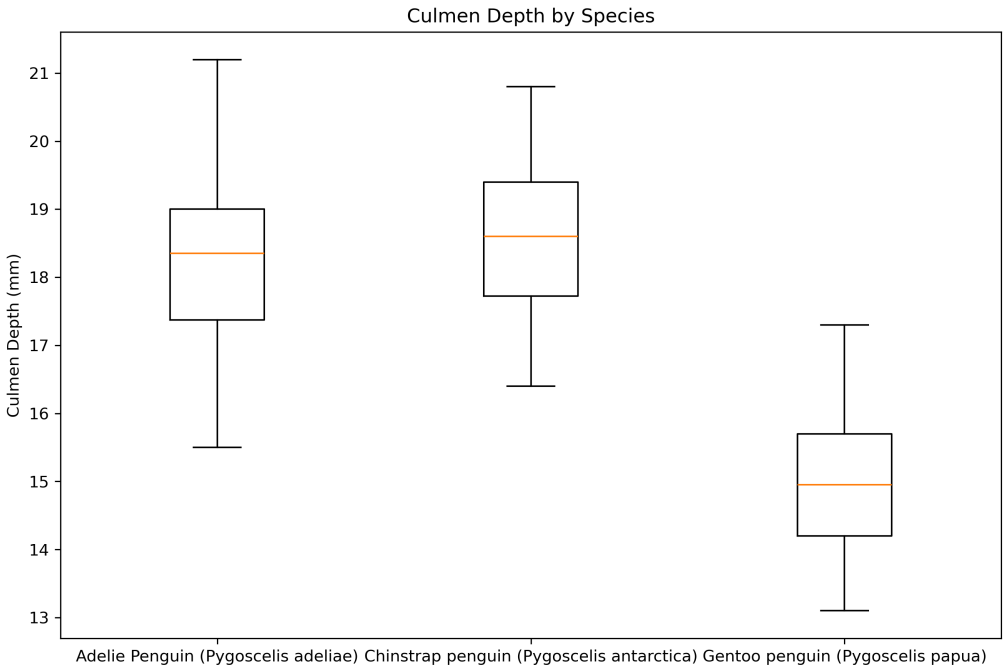

Figure 3: Box plot of culmen depth by species

The box plot of culmen depth by species shows that the Gentoo penguin has a much lower mean than the other two species. The box plots for the Chinstrap and Gentoo penguins are fairly symmetric which indicates that about as many penguins have the same variance from the median for penguins with shallow culmens as deep culmens.

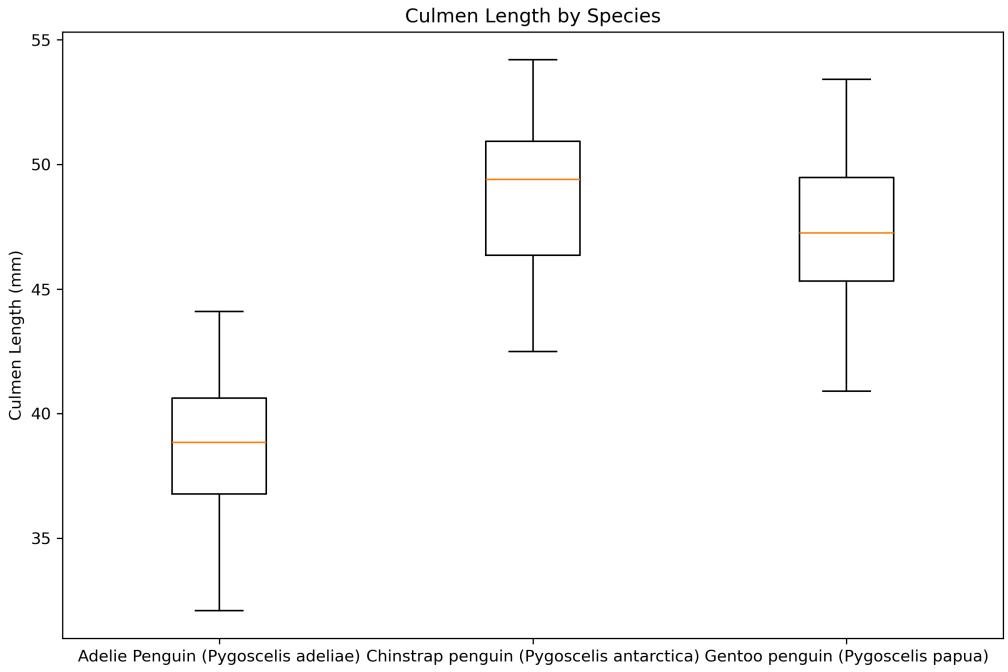

Figure 4: Box plot of culmen length by species

The box plot of culmen length by species shows that the Adelie penguins have shorter culmens than the Chinstraps and Gentoos. There are shape of each box plot are similar to the culmen depth box plots indicating an fairly even distribution within each species.

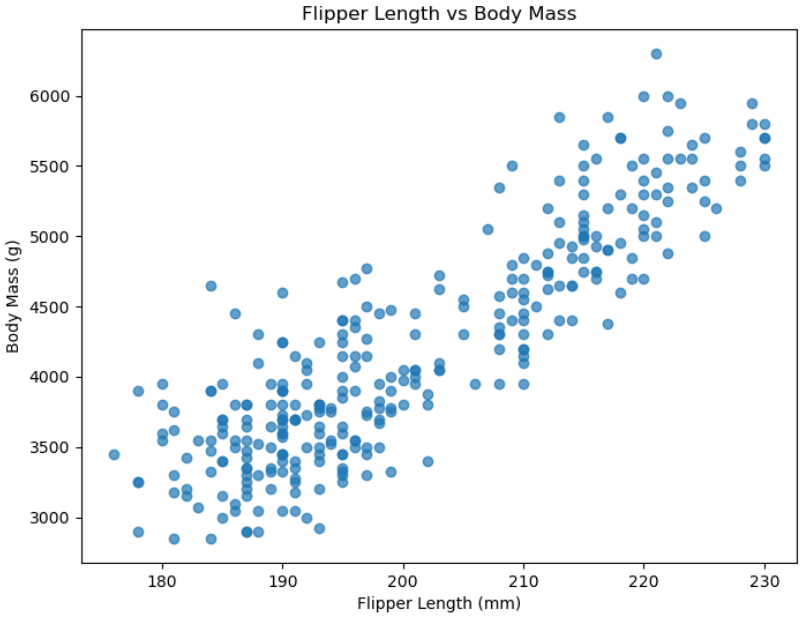

Figure 5: Scatterplot of body mass as a function of flipper length

The scatter plot of body mass as a function of flipper length shows a positive correlation between the two variables. The correlation appears to be strong, but more analysis is needed to determine the regression.

3 Regression Modeling

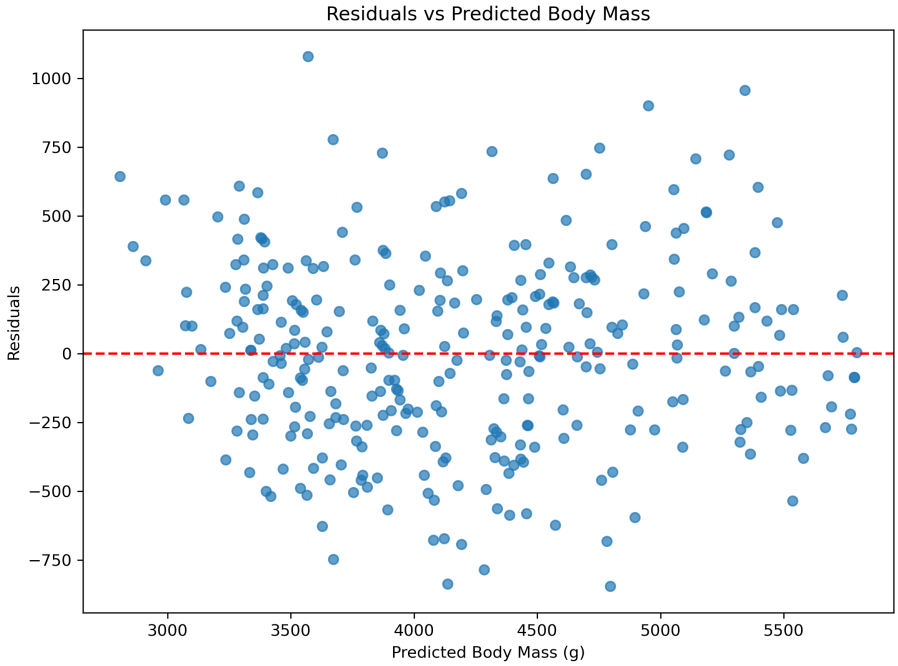

Figure 6: Plot of residuals compared to predicted body mass

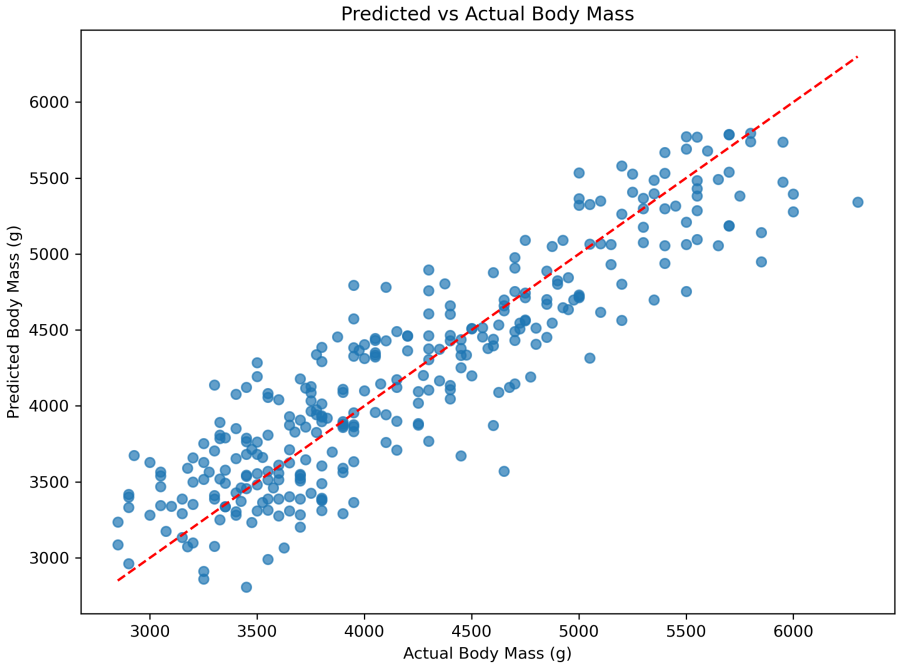

Figure 7: Plot of predicted body mass compared to data

Table 1: OLS Regression Results for Body Mass Prediction

Variable	Coefficient	Std. Error	t-statistic	P-value	95% CI
const	-5659.6232	293.858	-19.260	0.000	[-6237.861, -5081.386]
Culmen Length	-7.8206	5.304	-1.474	0.141	[-18.258, 2.616]
Flipper Length	49.8914	1.936	25.772	0.000	[46.082, 53.701]
Sex	361.1804	42.115	8.576	0.000	[278.308, 444.052]

The coefficients show that flipper length and sex both have correlation to body mass. This tracks with both intuition and the box plots of body mass and flipper length as compared to sex from the previous section. The R squared value is 0.812 so the three parameters explain 81 percent of the variance in body mass for the population.

4 Final Figures

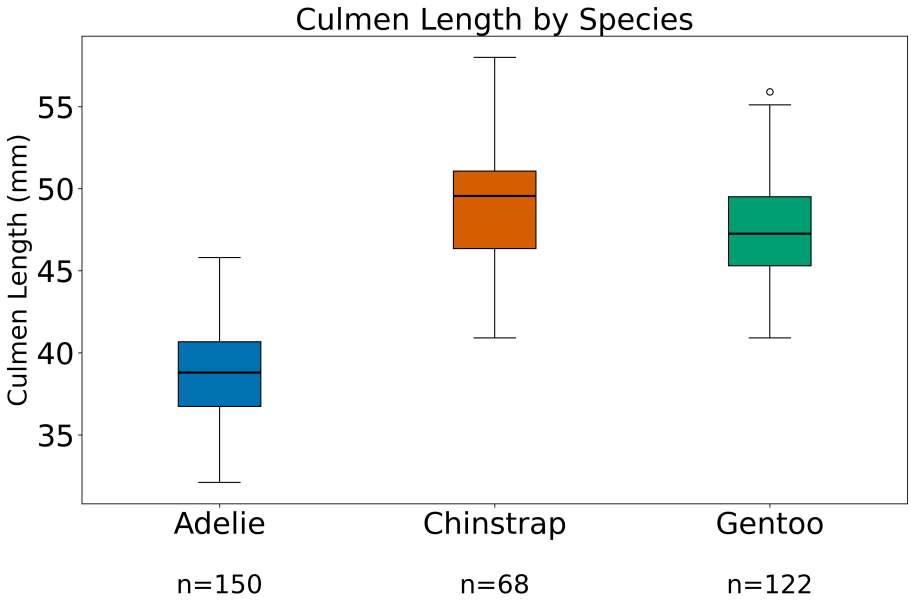

Figure 8: Box plot of culmen length for Adélie (Pygoscelis adeliae), Chinstrap (Pygoscelis antarcticus), and Gentoo (Pygoscelis papua)

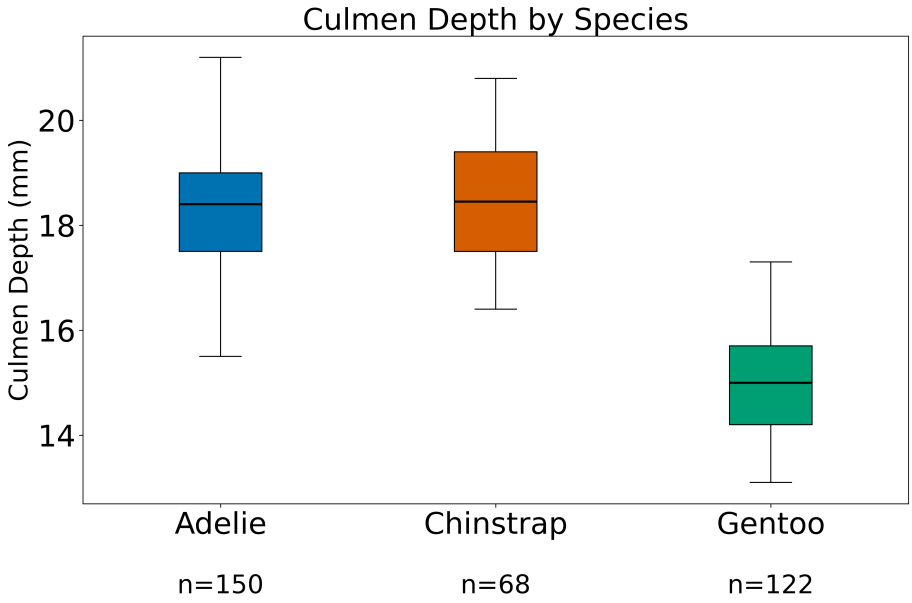

Figure 9: Box plot of culmen depth for Adélie (Pygoscelis adeliae), Chinstrap (Pygoscelis antarcticus), and Gentoo (Pygoscelis papua)

These figures were chosen as they tell a clear story. Since the dependent variables are measured against the same independent variables, it is easy to create matching color and symbolic schemes that guide the reader's eye. The data also shows clear differences in culmen shape between penguin species. This makes it easy for a reader to visualize the differences and similarities between the species. The box plots make it clear that Chinstrap penguins have the largest culmens of the species while the Gentoo have longer but less deep culmens and Adelie have shorter and deeper culmens. As stated before, the general symmetry around the median suggests that the middle of the data is evenly distributed about the median. For maximum readability, I wanted to keep the figures as simple as possible. This meant removing the scientific names from the plots and putting them in the captions, color coordinating the boxes between plots, and providing sample sizes for each box. When color coordinating, I ensured that the colors were colorblind friendly. Python generated png files were used with no post processing. I think there could be a little bit less white space, and maybe the outliers should have been included as it is a box plot that is designed to show the spread of data in the form of inter-quartile ranges.