Hyperiondev

TASK

Exploratory Data Analysis on the Penguins_lter Dataset

Visit our website

Introduction

The penguins data set contains data on size measurements, clutch completion status, sexes, blood isotopes, and locations for three species of penguin: Adélie, Chinstrap, and Gentoo.

DATA CLEANING

The columns were investigated by looking at their unique values (*peng_df['column name'].unique()*) and it was determined that 'studyName', 'Species', 'Island', 'Clutch Completion', 'Date Egg', 'Culmen Length (mm)', 'Culmen Depth (mm)', 'Flipper Length (mm)', 'Body Mass (g), and 'Sex' would be of most use. All other columns were removed.

The 'Comments' column was mostly empty, but contained some information as to why data might be missing. Most data was missing completely at random. This column was then removed as it would not be useful for the data analysis.

One value in the 'Sex' column was '.', so this was changed to match the NaN values.

The 'Date Egg' column was converted to datetime format.

Breakdown of Column Information

studyName – Sampling expedition from which data were collected, generated etc. There are 3 different studies in the dataset: 'PAL0708', 'PAL0809', and 'PAL0910'

Island – A character string denoting the island near Palmer Station where samples were collected. Penguins can be from the island of Torgersen, Biscoe, or Dream.

Clutch Completion – A character string denoting if the study nest observed with a full clutch, i.e., 2 eggs. Can be 'Yes' or 'No'.

Date Egg – A date denoting the date study nest observed with 1 egg (sampled)

Culmen Length (mm) – A number denoting the length of the dorsal ridge of a bird's bill (millimeters)

Culmen Depth (mm) – A number denoting the depth of the dorsal ridge of a bird's bill (millimeters)

Flipper Length (mm) – A number denoting the length penguin flipper (millimeters)

Body Mass (g) – A number denoting the penguin body mass (grams)

Sex – Birds are categorised as 'FEMALE' or 'MALE'

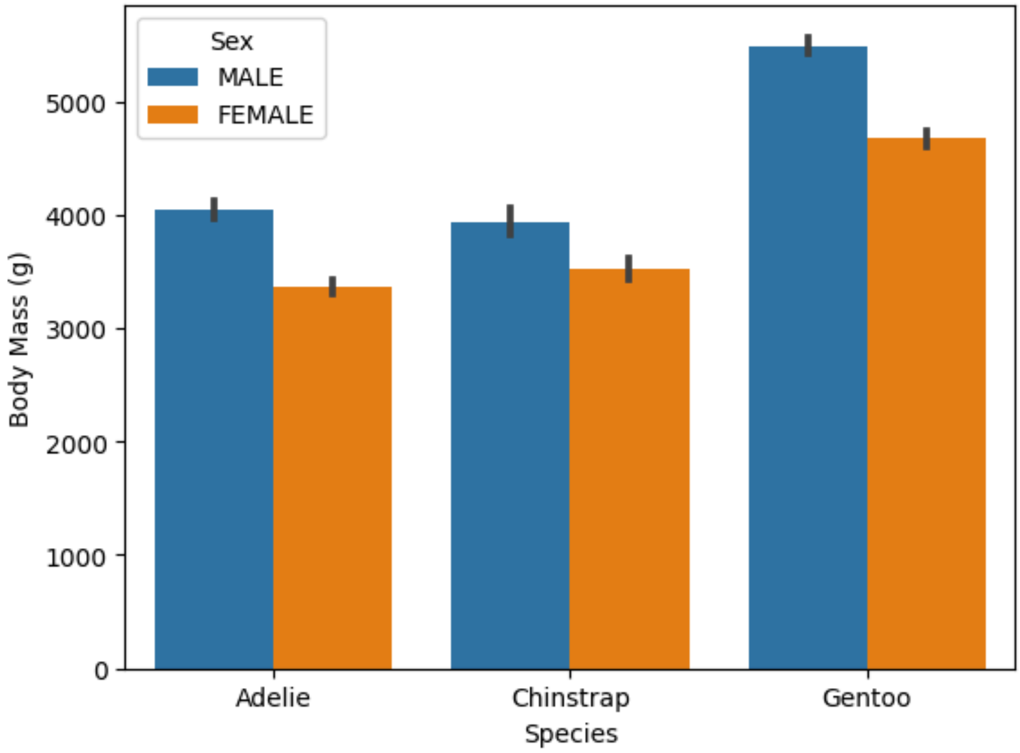
MISSING DATA

Missing data in the chosen columns was as follows:

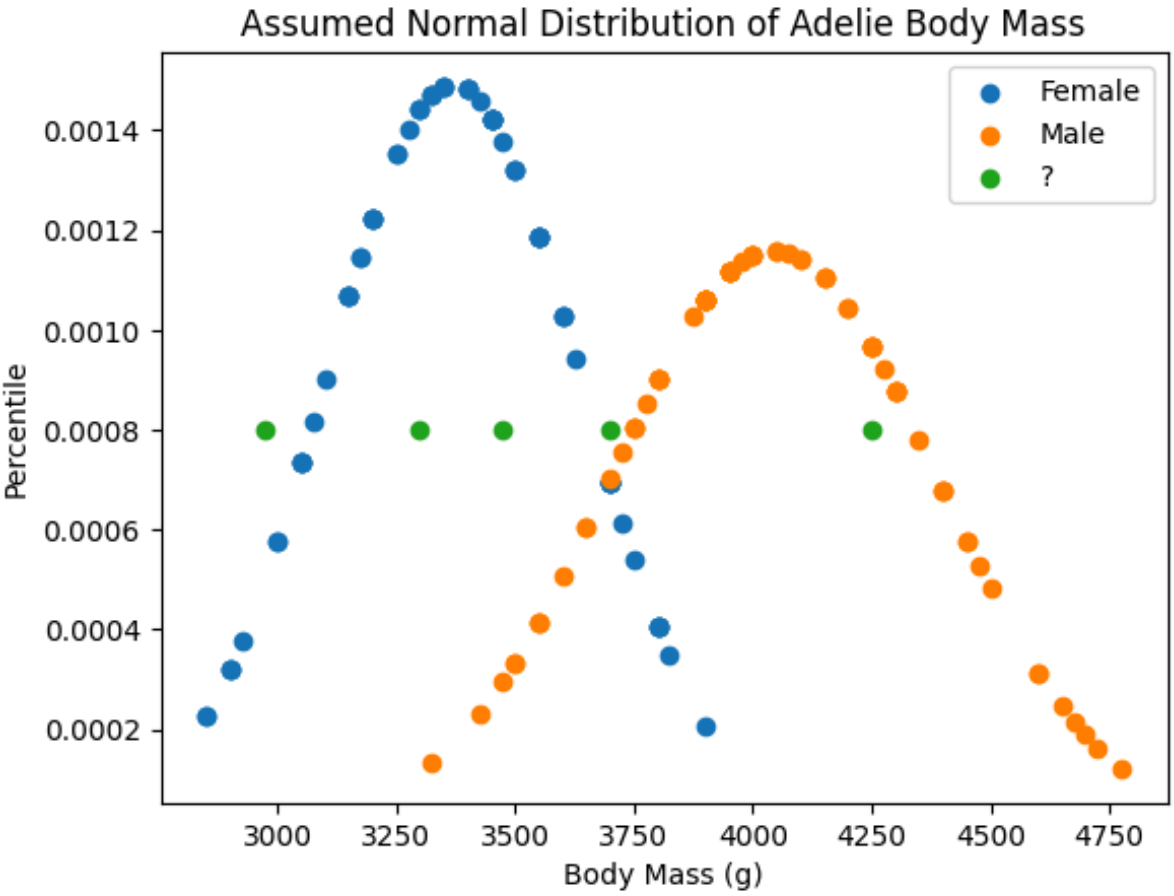
Column Name	No. of NaN Values
studyName	0
Species	0
Island	0
Clutch Completion	0
Date Egg	0
Culmen Length (mm)	2
Culmen Depth (mm)	2
Flipper Length (mm)	2
Body Mass (g)	2
Sex	11

For columns with continuous data, the missing values were replaced with the mean value for the column.

For the 'Sex' column, it was determined that body mass could be used as an indicator of sex.



The penguins with missing sex data belonged to the Gentoo and Adélie species. To make an accurate estimation of their sex, a visualisation would be required that showed approximately how distant a penguin was from the average weight for males and females of their species.



Judging by the above graph:

ADELIE penguin weighing

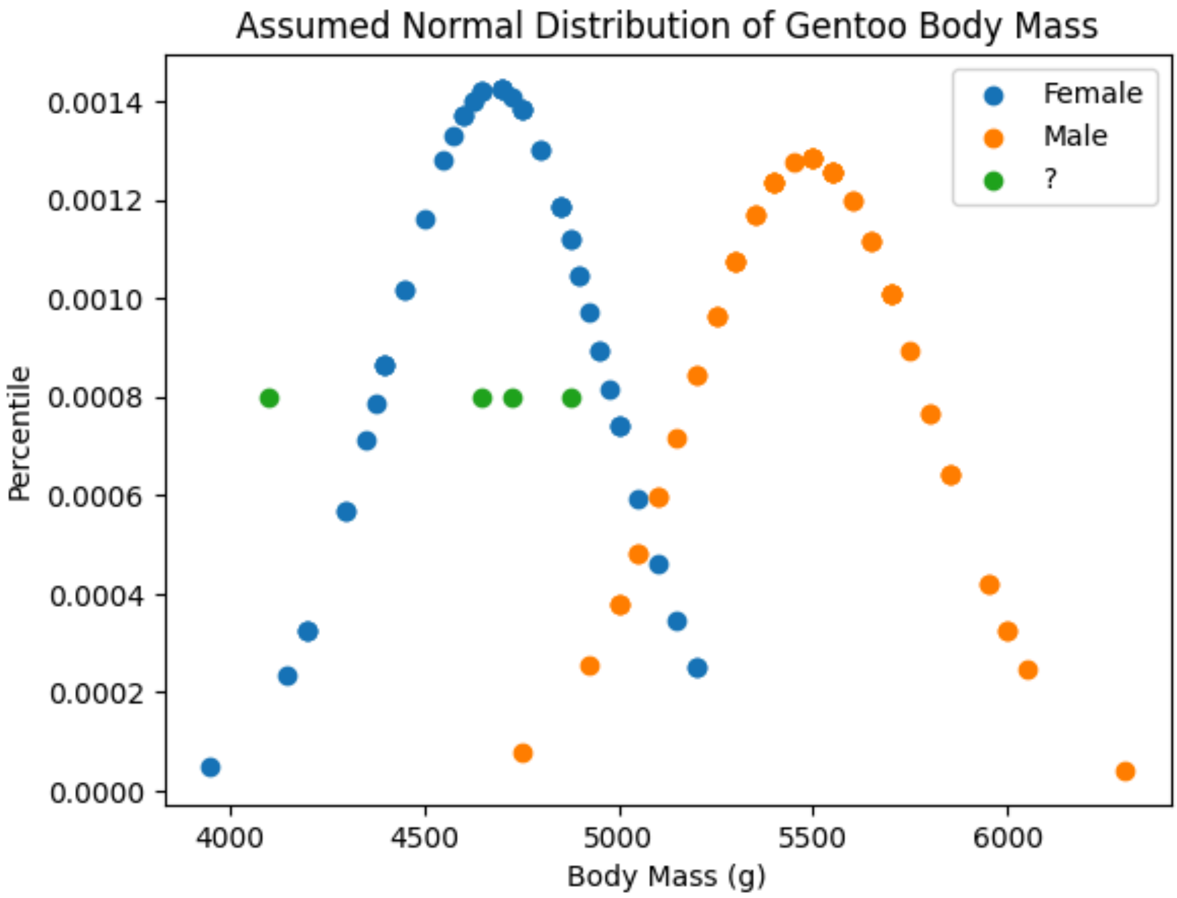
4250g: Extremely likely to be MALE

3700g: Slightly more likely to be MALE

3475g: Very likely to be FEMALE

3300g: Very to be FEMALE

2975g: Very likely to be FEMALE



Judging by the above graph:

GENTOO penguin weighing

4875g: Likely to be FEMALE

4725g: Very likely to be FEMALE

4650g: Very likely to be FEMALE

4100g: Extremely likely to be FEMALE

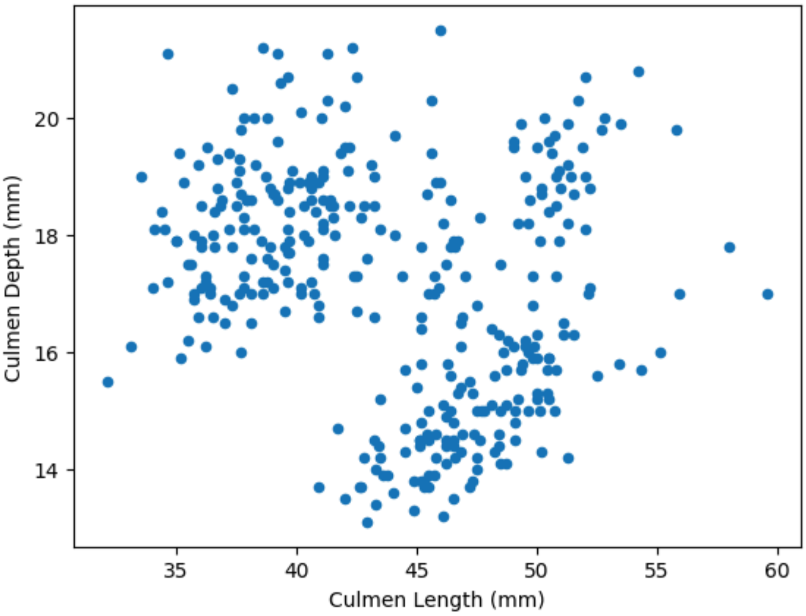
Two penguins remained unidentified: the penguin at index 339, and the penguin at index 3 as their masses had been missing and filled in by the column average.
Both turned out to be the source of the missing data in the other body measurement columns.

As the majority of their rows were made up of averages and there was no way to accurately guess their sexes, their rows were removed.

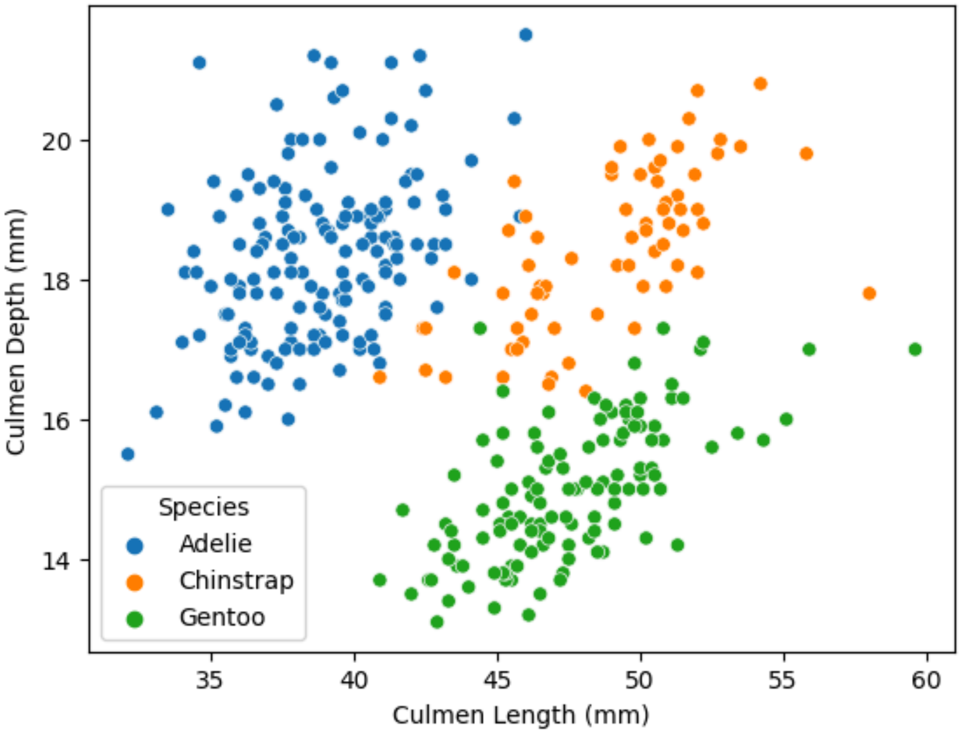
All missing data had now been dealt with.

DATA STORIES AND VISUALISATIONS

First a graph was created to show correlation between bill length and depth.



Although the graph initially appeared fairly meaningless, it became clear that dividing the dataset by species would be necessary for accurate exploration.

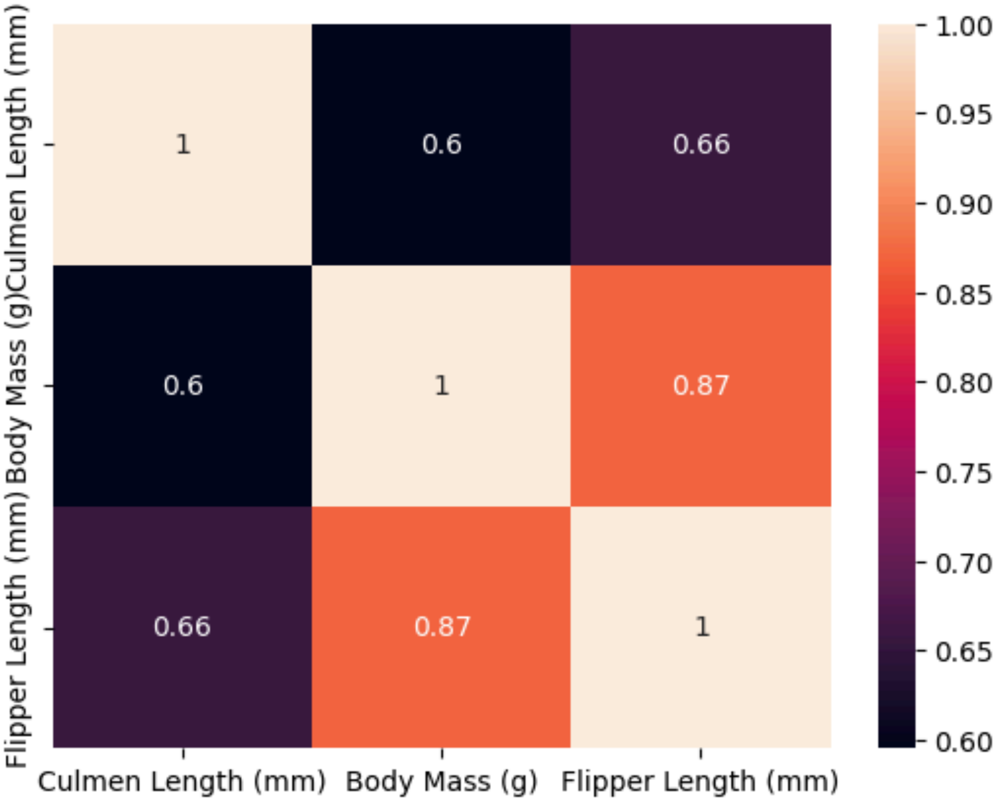


The same graph now divided by species shows a clear positive correlation between bill (‘culmen’) length and depth.

Adélie Penguins tend to have shorter but deeper bills, Gentoo Penguins have longer but thinner bills, and Chinstrap Penguins have bills that are on average slightly longer than the Gentoos, and comparably deep as the Adélies.

The clustering shows the three species are strongly defined by their bill lengths and depths.

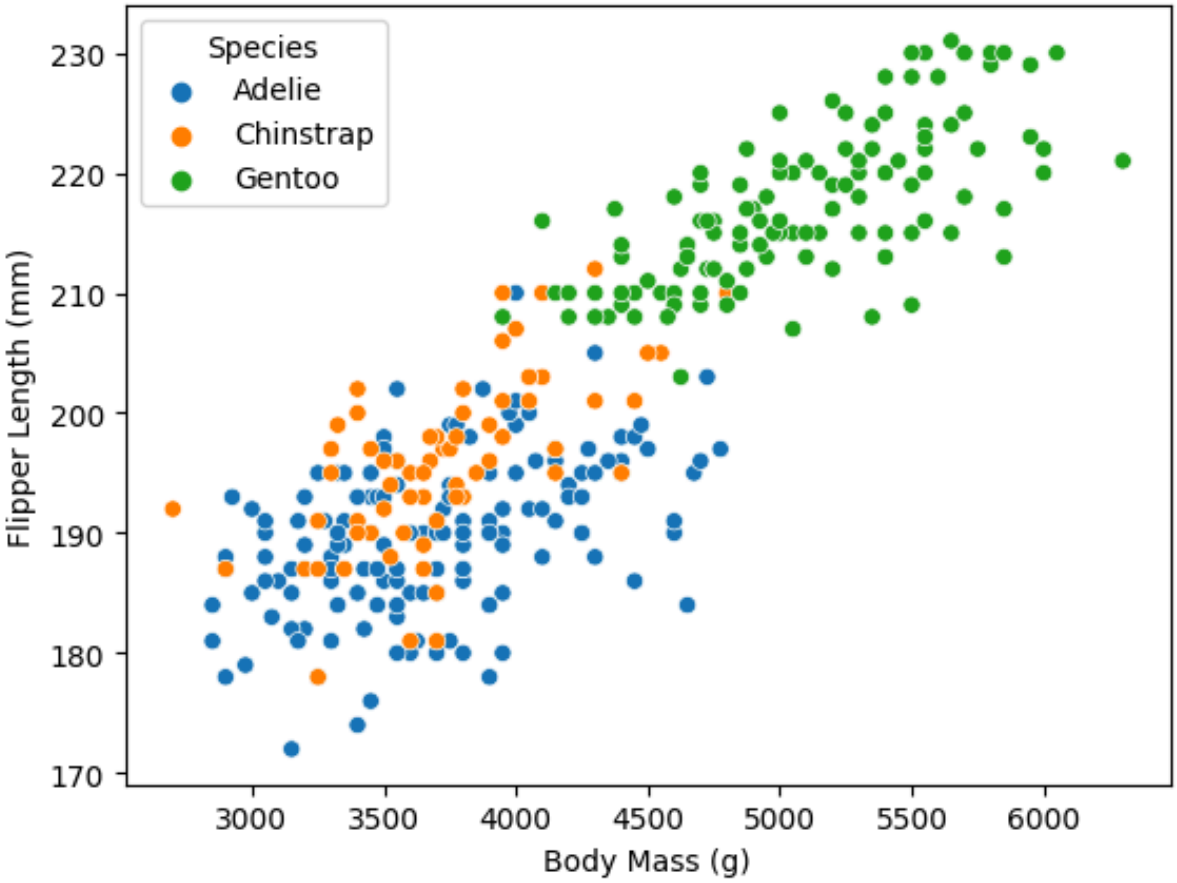
Correlation between Flipper Length, Body Mass, and Culmen Length



This graph shows that there is a very strong correlation between Body Mass and Flipper Length. While there is still a correlation between Body Mass and Culmen Length, it is not as strong.

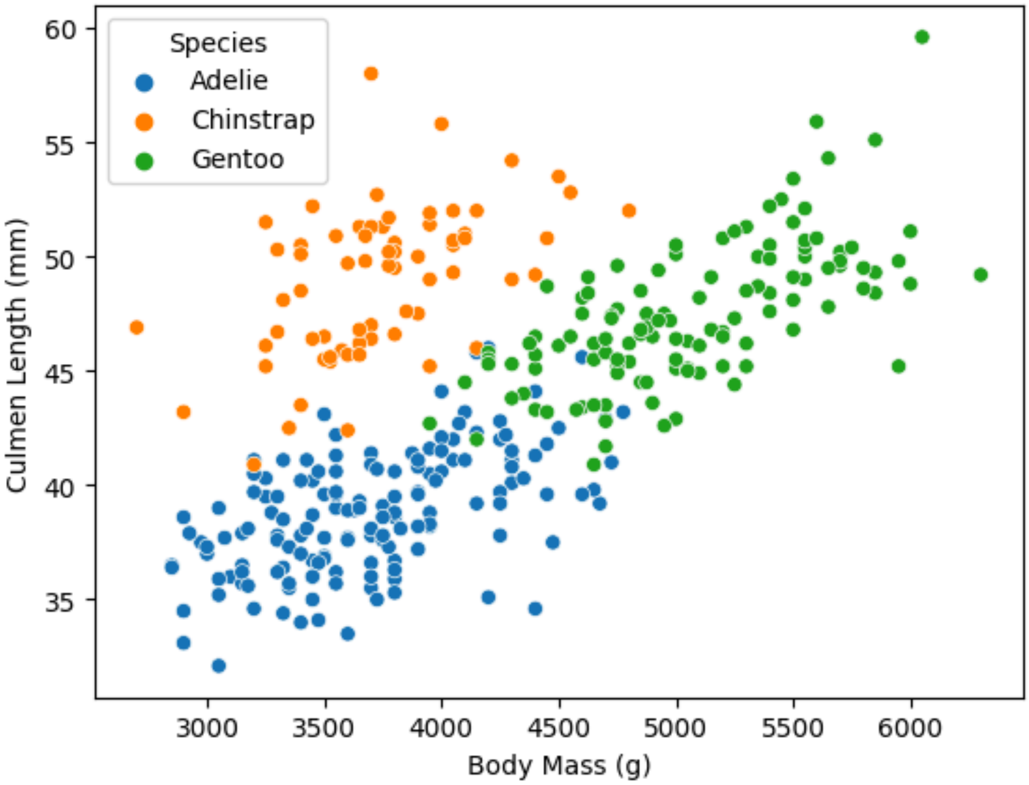
This may be because heavier penguins need longer flippers for propelling themselves through the water, however, as their beak function is not mobility related (eg. catching fish), they wouldn't have a need for beaks to increase in size as proportionally to their body mass (eg. the fish are still the same size so they do not require a larger beak to catch them).

These results were then double checked against species.



The graph shows a strong correlation between Flipper Length and Body Mass across species, upholding the reading of the heat map.

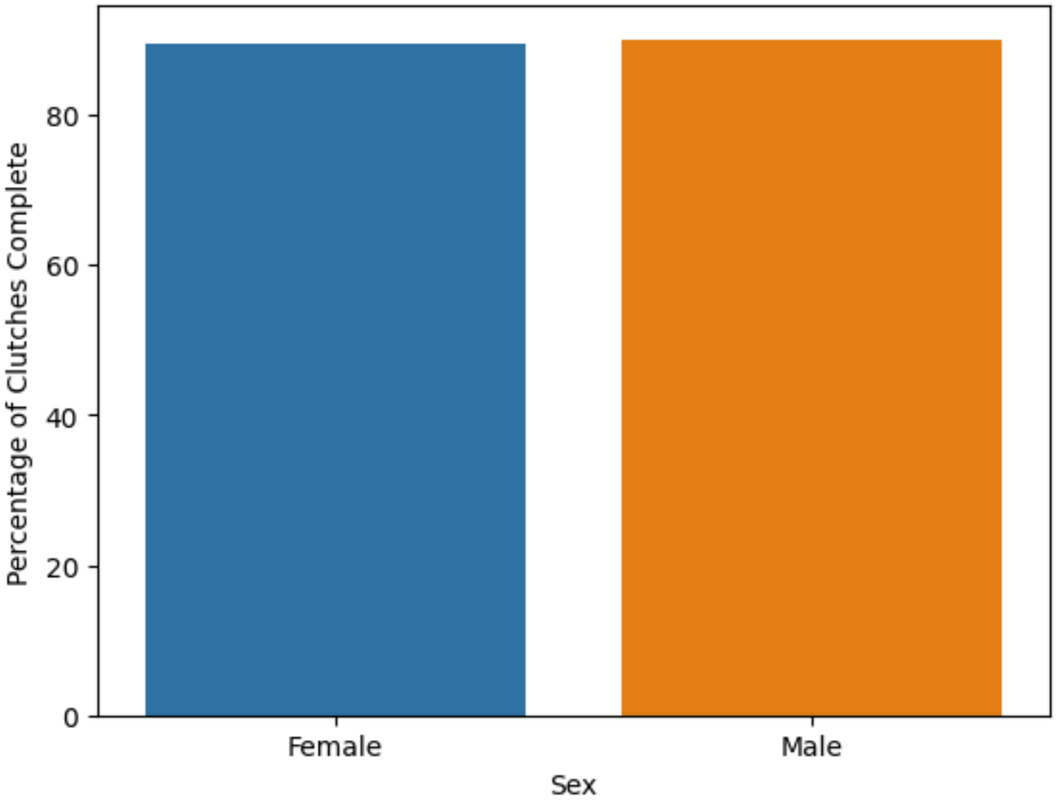
The graph also shows that Gentoo are the larger of the penguin species and have longer flippers. Chinstraps have longer flippers for their size than Adélies, but have about the same body weight range as them.



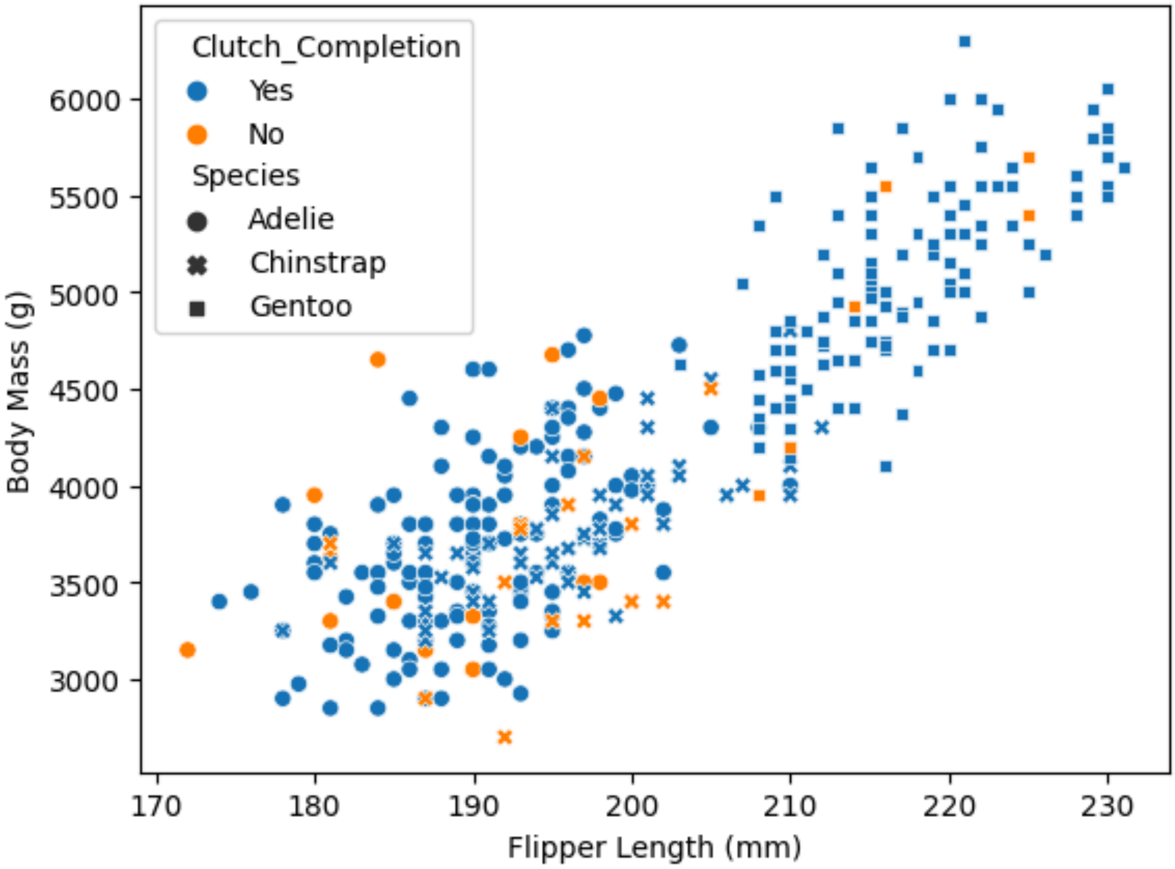
The above graph also supports the conclusions drawn from the heat map, showing that while there is some correlation between bill length and body mass, it is much weaker than with flipper length, and there are many more outliers.

Predicting the Likelihood of a Clutch being Complete

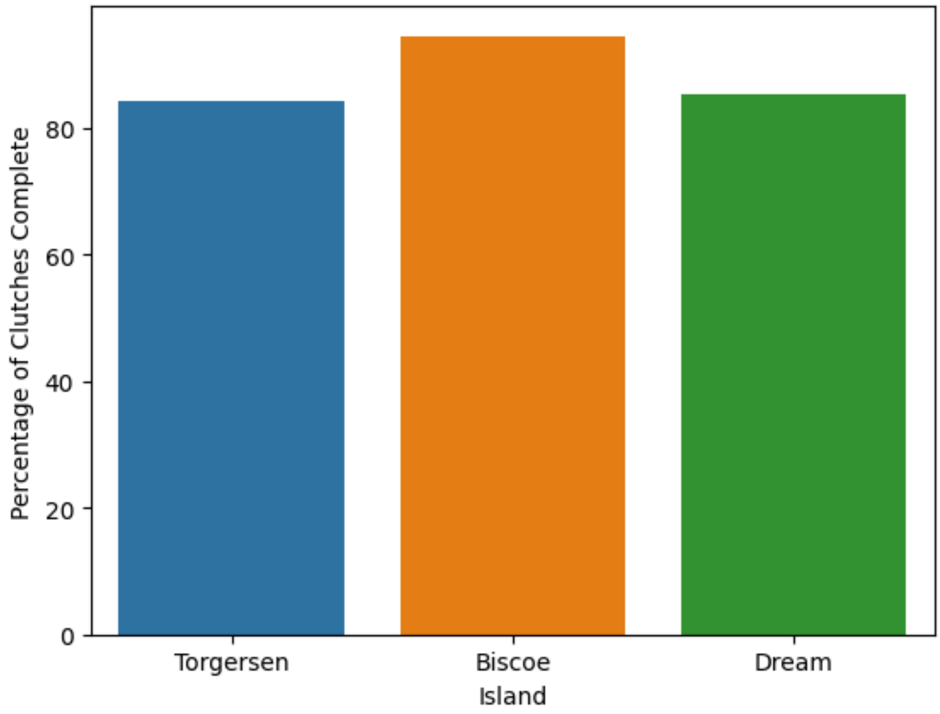
It was assumed that sex would not affect clutch completeness as each clutch likely has one female and one male parent. A graph was made to check this assumption was accurate.



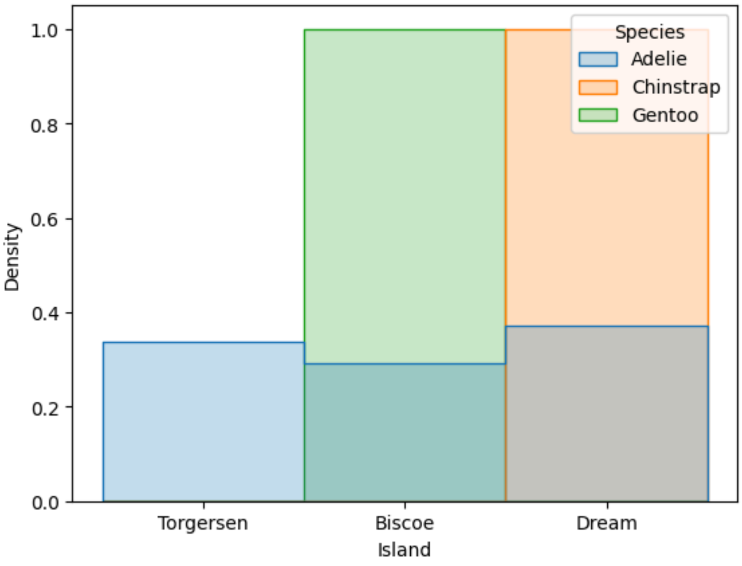
Next, an investigation was done to ascertain whether body mass and clutch completeness had any correlation, this necessarily had to also be compared with species.

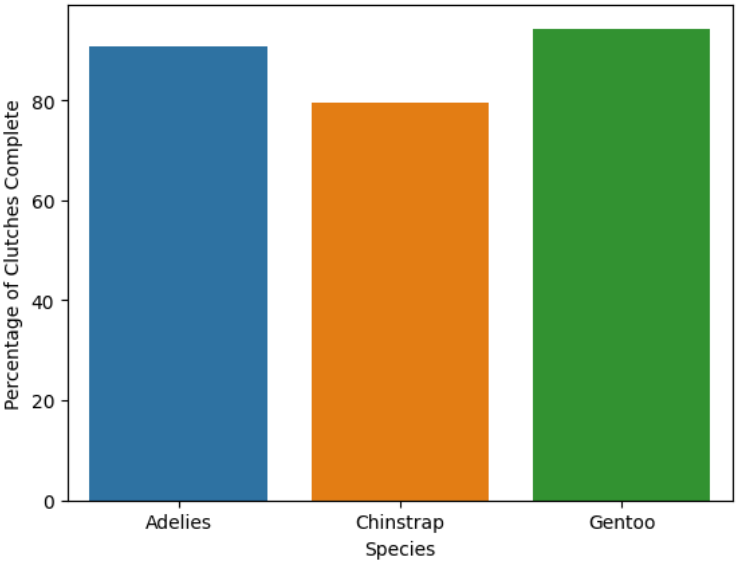


The graph shows no clear correlation the size of the parents and clutch completeness.



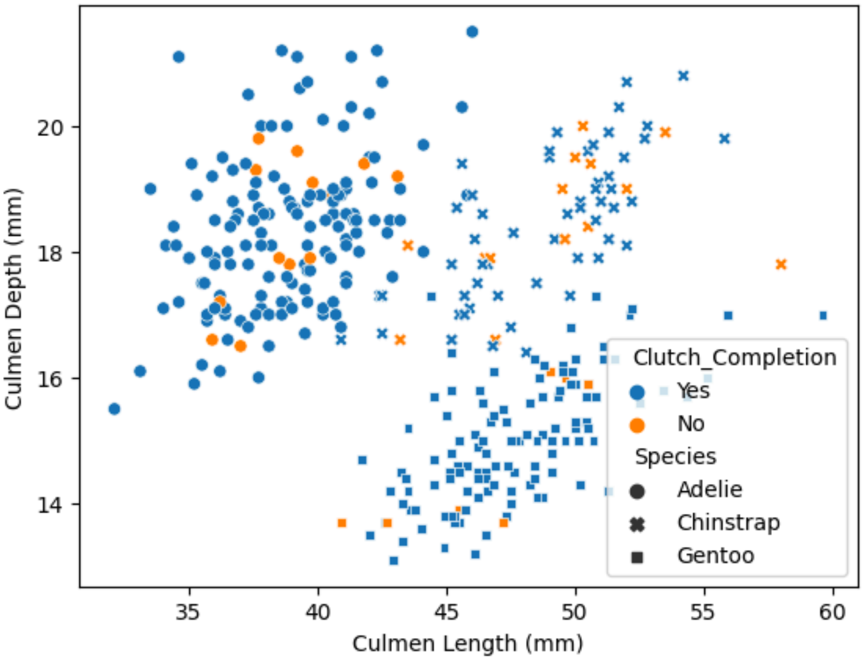
The island of Biscoe has slightly higher clutch completion rates than the other two islands.





We can see that Biscoe Island has the whole Gentoo population and none of the Chinstrap population, which resides wholly on Dream. It follows that the island has a slightly higher clutch completion, as the Chinstrap species has the lowest clutch completion and the Gentoo species has the highest.

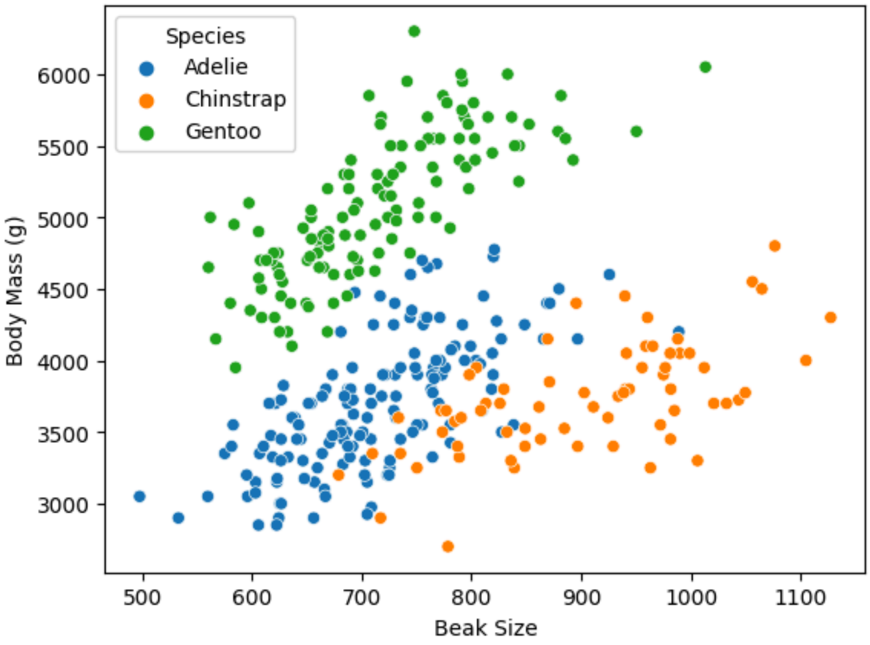
The Adélie species can be found in more or less equal density on every island.



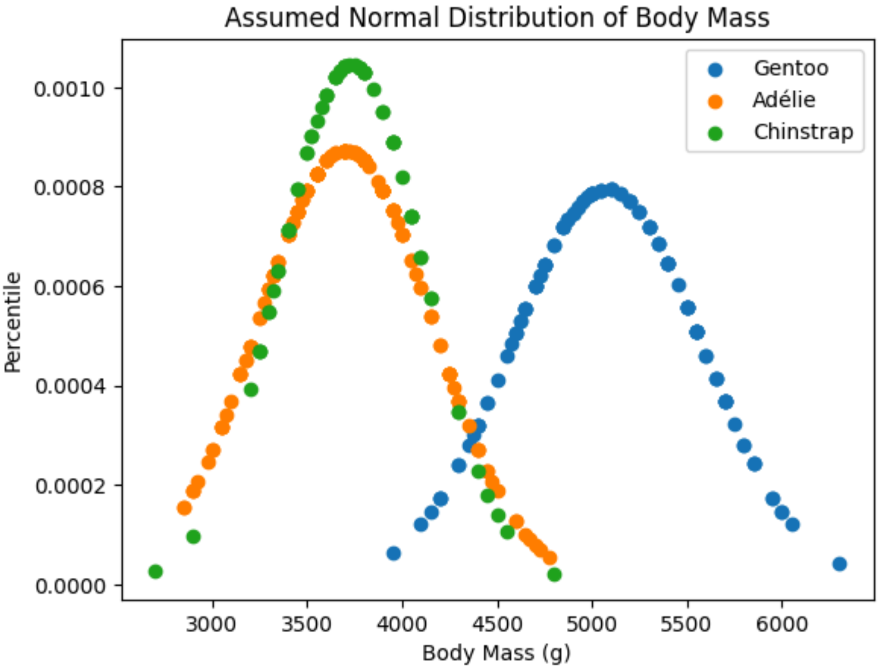
Overall, there appears to be no single strong predictor of clutch completion based on the measurements of the penguin or its location.

Comparing Beak Size and Body Mass

A ‘Beak Size’ column was created by multiplying the values in the ‘Culmen Length’ column with those in the ‘Culmen Depth’ column in order to be able to compare both measurements at once.



We can see that despite being the smaller of the species, the Chinstraps have the overall largest beaks, though also the widest spread of beak size. Adelie are the smallest species in both body mass and beak size. While there is overlap in mass between the Chinstraps and Adélies, there is very little between the Gentoos and other species, as can be seen below.



THIS REPORT WAS WRITTEN BY : ROSIE YOUNG