Exploratory Data Analysis on the Penguin Size Dataset

Introduction

This dataset is related to penguin research and could be used for various data analysis or machine learning tasks, such as predicting a penguin's species based on its physical characteristics or investigating correlations between the different characteristics. The dataset has information on the Penguins species, habitat, gender, and information about its size. In this report I have conducted an exploratory data analysis on the dataset and provided a description of my findings accompanied by visualisations.

MISSING DATA

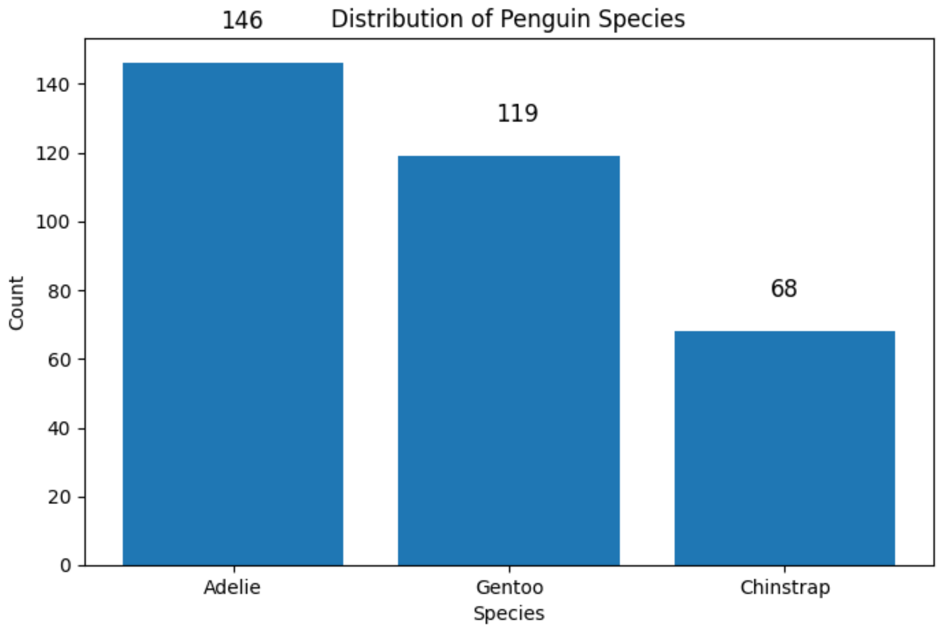
On inspecting the dataset at first glance, it appeared that some rows had missing data. Some of the missing data was represented as 'NA' while one of the data entries in the 'sex' column had been entered as '.' (a full stop). It was assumed that observations with incomplete data were entered as 'NA' during the data collection process, and assumed that the full stop entry in the 'sex' column was an accidental typing error while typing the data into the dataset. Out of the 344 observations, only a few had missing values, and so any rows from the dataset with missing values were removed. After this, the updated dataset had 333 rows of data, which meant that only 3.2% of the observations were removed due to them having missing or incorrect data values. This did not significantly change the overall meaningfulness of the data as it was only a very small fraction of the dataset that was removed.

DATA CLEANING

To start cleaning the data, first any duplicate rows were removed from the dataset as these can lead to a bias analysis or misleading visualisations. The data type of the data in the 'Flipper Length' and 'Body Mass' columns were converted from float64 to int64. This is because the data in these columns were all whole numbers and int64 typically uses less memory compared to float64. This also helps to improve the performance and speed of the data manipulation and analysis.

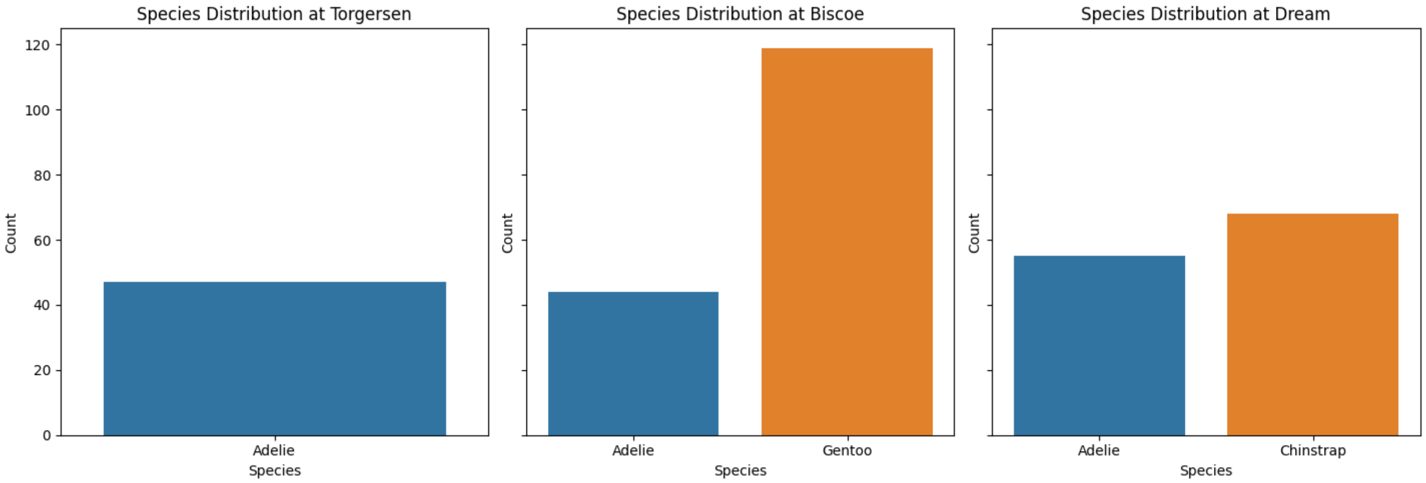
DATA STORIES AND VISUALISATIONS

To better understand the distribution of Penguin species in the dataset a bar chart was created which provides a clear visualisation of how many observations were taken for each penguin species during the data collection process.

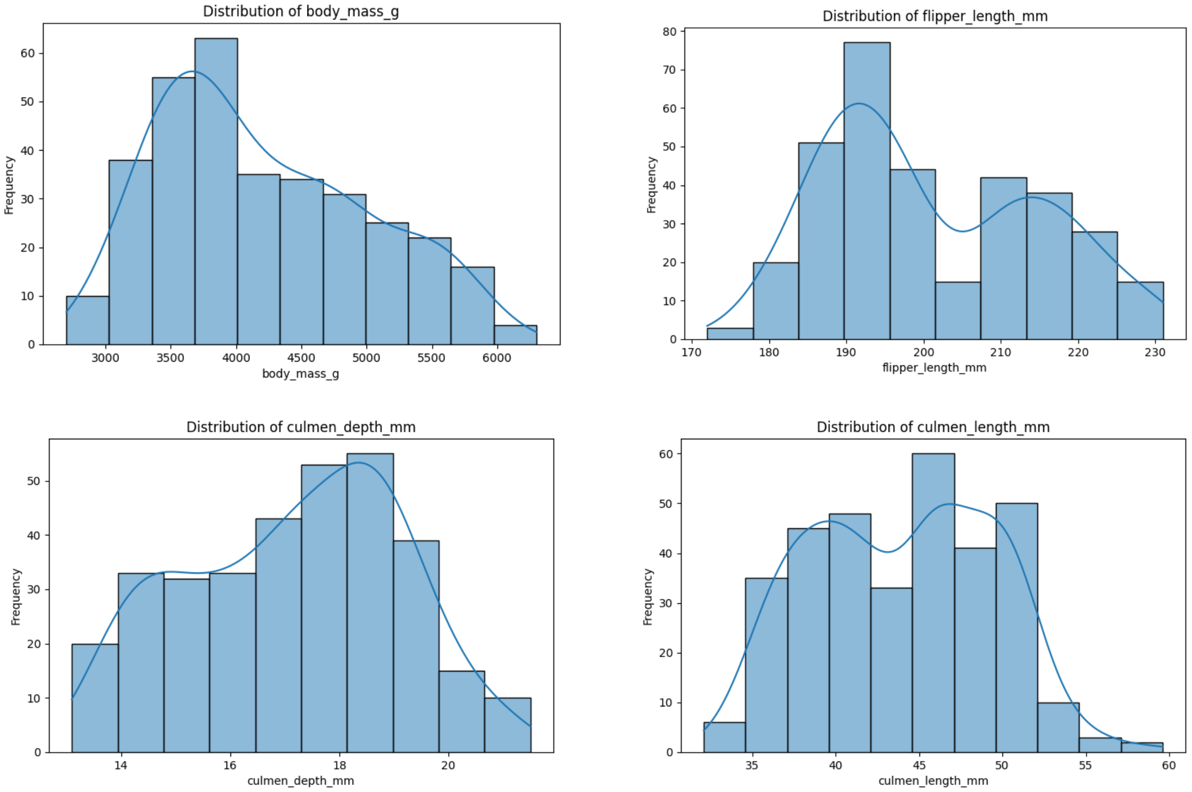


The bar graph above shows that the Adelie species was the most frequently recorded penguin, with 146 observations, followed by the Gentoo species with 119 observations. The Chinstrap species was the least frequently recorded penguin, with only 68 observations. This is less than half as many as the Adelie species. The data indicates that the Adelie species is the most abundant species of Penguin in the area where the data was collected, while the Chinstrap species are the most sparse.

The distribution of penguin species across the three different islands (Torgersen, Biscoe and Dream islands) was then inspected. This gave a good insight into which species inhabit which island, and explained why the Adelie species is so abundant. As seen in the bar graphs below, the Adelie species inhabit all 3 of the islands while the Gentoo and Chinstrap penguins only inhabit one island each. The Gentoo inhabits the Biscoe Island while the Chinstrap inhabits the Dream Island. This is useful data for understanding what habitats are most suitable for each species of penguin.

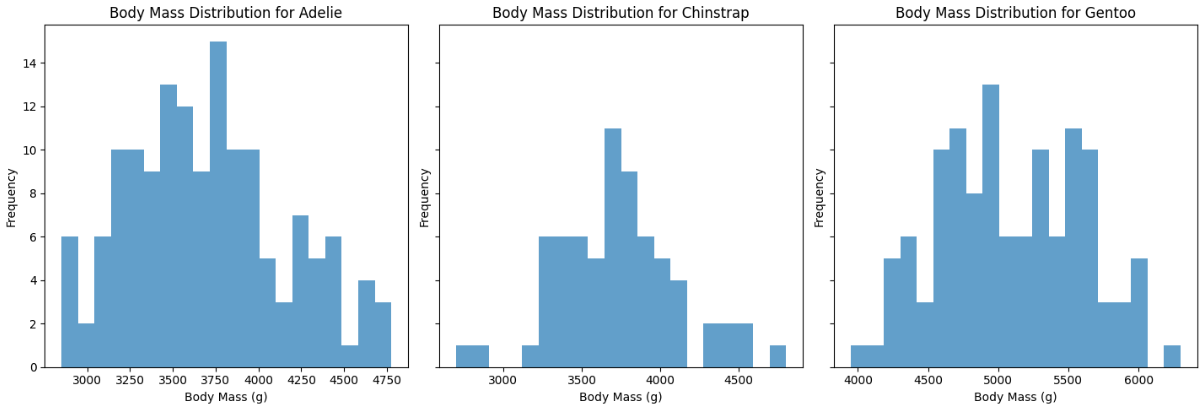


Delving deeper into the size measurements of the penguins flippers, culmens and body mass, we can use the histograms below to show the distribution of each measurement across all the penguins in the dataset. These give an overview of the data for all the penguins that were observed. These histograms show us that the most common body mass of the penguins was around 3.75-4kg, the most common flipper length was around 190-195mm, the most common culmen depth was around 18-19mm, and the most common culmen length was around 45-47mm.



What was even more insightful was looking at each species separately to compare them. This provided a deeper understanding of the nuances and differences between the three species of penguins present in the dataset.

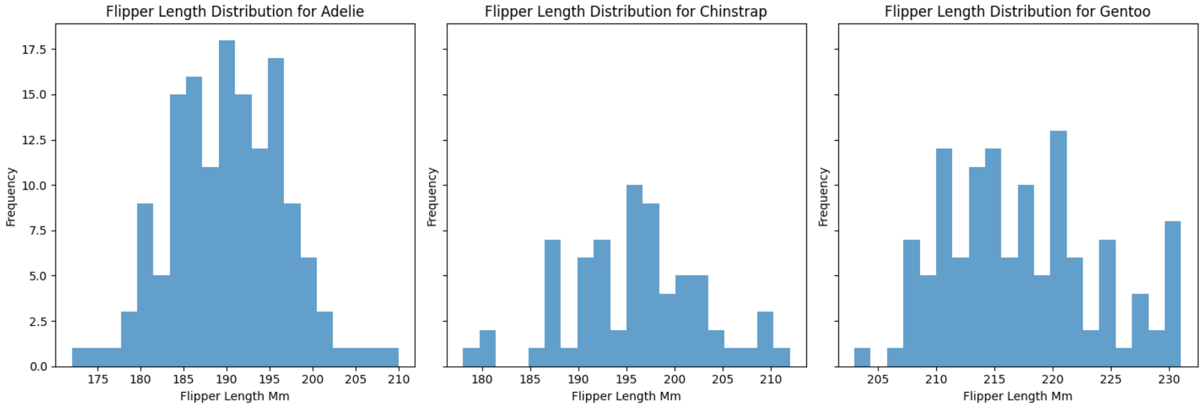
We can see from the size distribution histograms on the next page, for example, that the body mass of the Chinstrap species has less variance than the other two species. The Chinstrap species has a normal distribution for body mass measurements, however this is not the case for the species culmen depth or culmen length which have more variance and are less normally distributed. In simpler terms this means that the culmen (the upper ridge of the penguins beak) of the Chinstrap species comes in many different sizes while the culmen of the other two species are less variable.



Body Mass Distribution for Adelie

Body Mass Distribution for Chinstrap

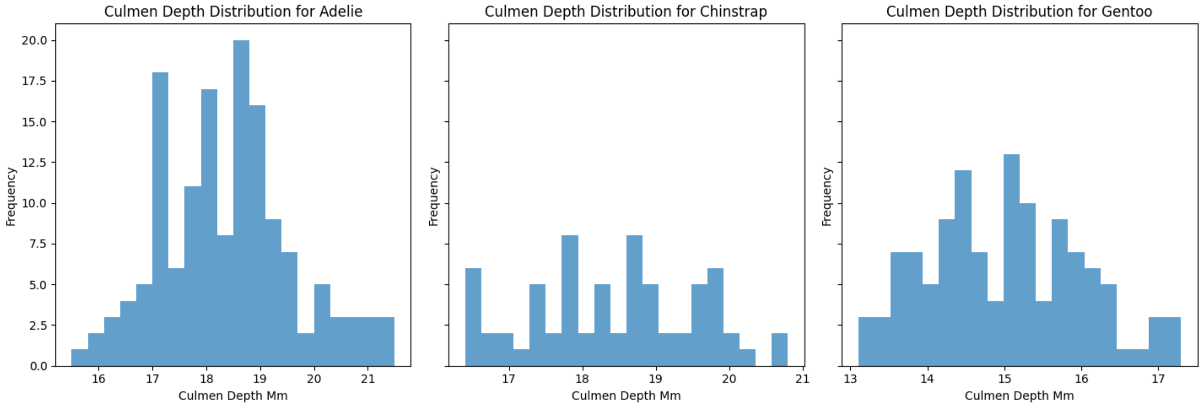
Body Mass Distribution for Gentoo



Flipper Length Distribution for Adelie

Flipper Length Distribution for Chinstrap

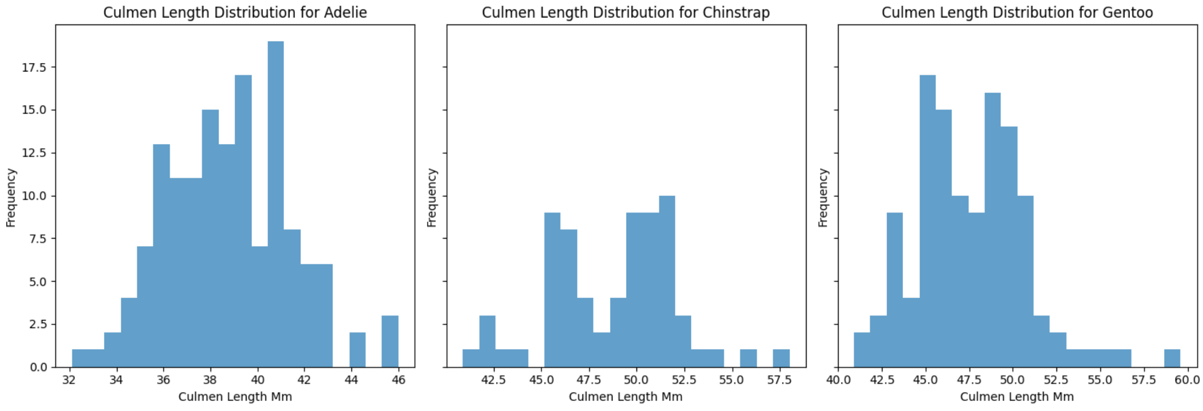
Flipper Length Distribution for Gentoo



Culmen Depth Distribution for Adelie

Culmen Depth Distribution for Chinstrap

Culmen Depth Distribution for Gentoo



Culmen Length Distribution for Adelie

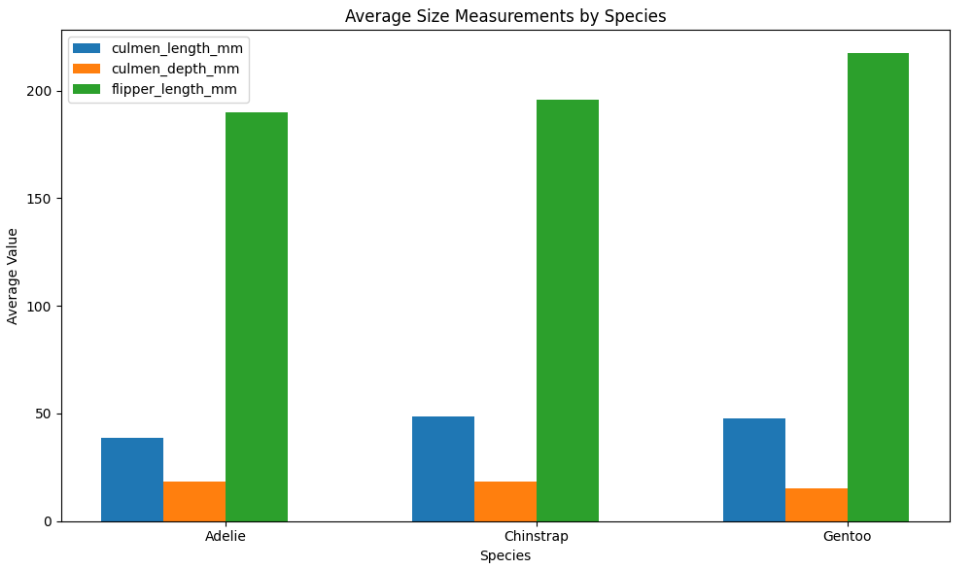
Culmen Length Distribution for Chinstrap

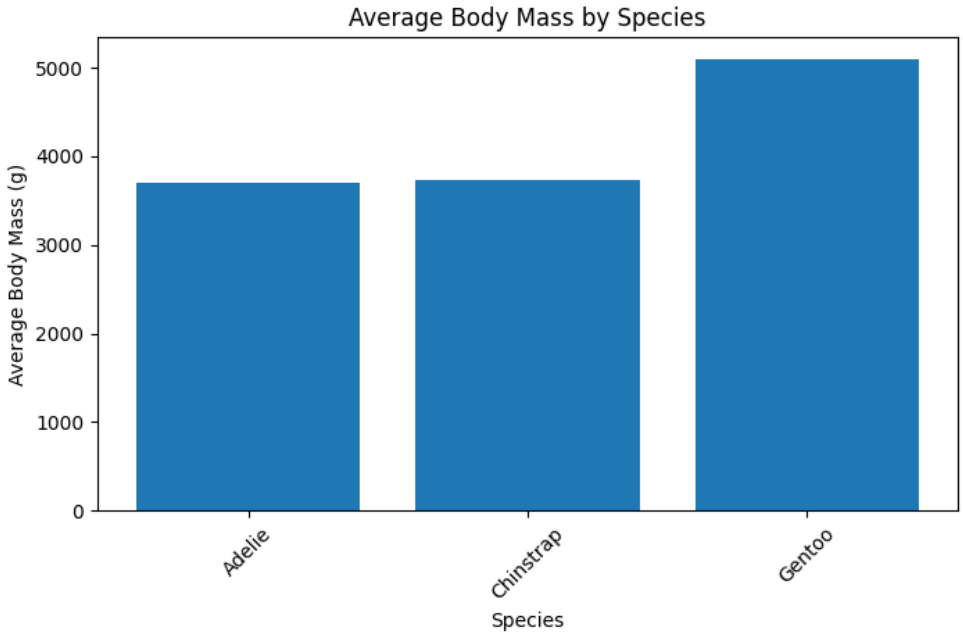
Culmen Length Distribution for Gentoo

Reading further into the histograms on the previous page, we can see that the most common body mass for the Adelie and Chinstrap species is around 3.75kg, while the Gentoo species is much higher at around 4.9kg.

The most common culmen depth for the Adelie species appears to be deeper than the most common culmen depth for the Gentoo species, however the most common culmen length for the Adelie species appears to be shorter than the most common culmen length for the Gentoo species. This suggests that in general the culmen of the Adelie species is deeper but shorter than the culmen on the Gentoo species.

The flipper length of the Gentoo species is not normally distributed but the most common length is around 221mm, while the Adelie species is 191mm and the Chinstrap species is 196. It makes sense that the Gentoo species would have larger flipper length because they have a higher average body mass than the other two species. The average size measurements for each species can be seen in the two bar graphs below, where culmen length, culmen depth and flipper length are shown on the first graph and body mass are shown on the second graph.

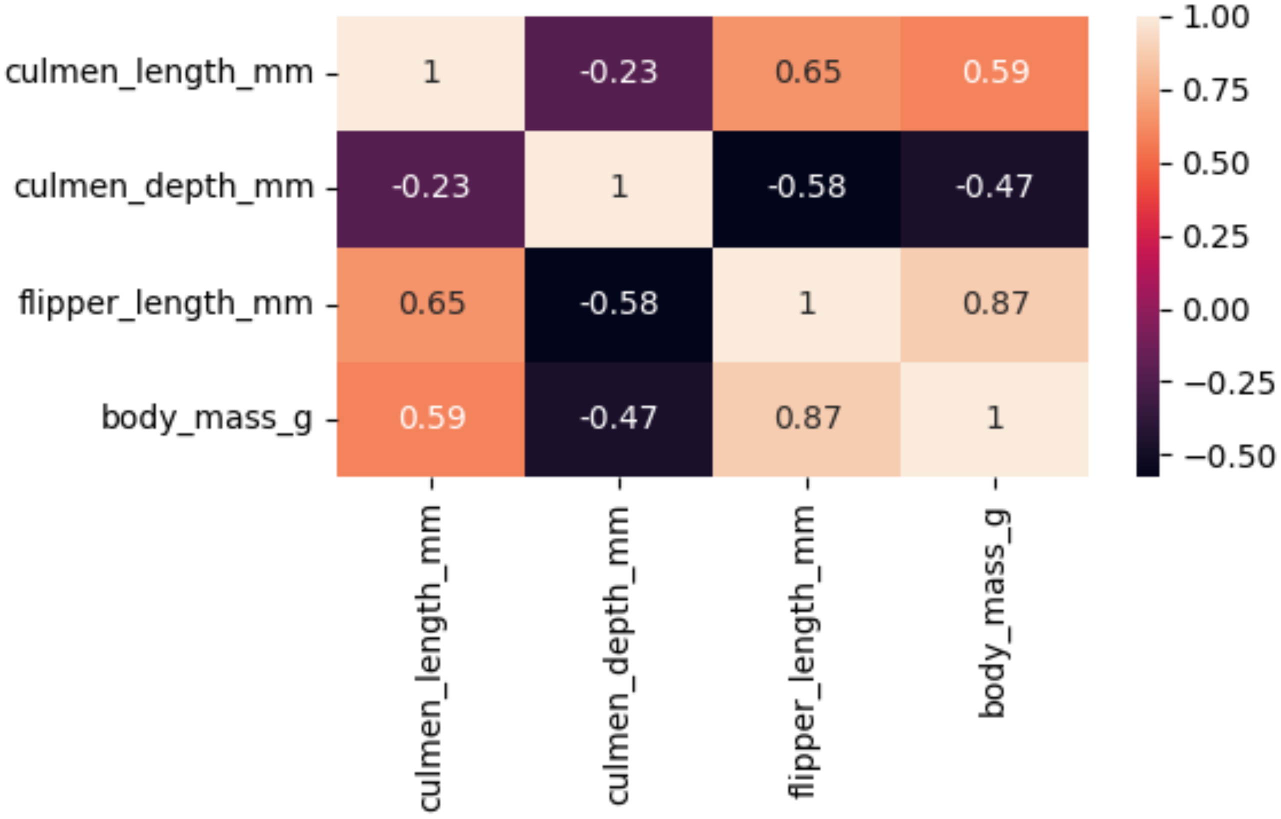




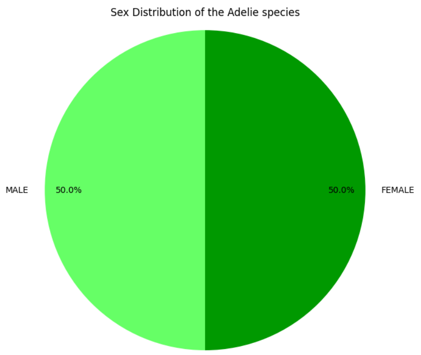
A heat map was created to visualise the correlations between the measurements across all the penguins. If a more in depth analysis was to be done on the difference between each species this type of heat map could be created for the size measurements of each species.

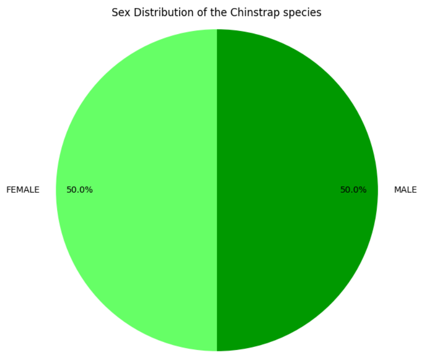
But for the purpose of this report, the heat map below indicates a strong positive correlation of 0.87 between flipper length and body mass. This coincides with our earlier discovery that the Gentoo species has the largest body mass and the largest flipper length. There is also a positive correlation of 0.65 between flipper length and culmen length. This means that in general, the longer the flipper is of a given penguin, the larger the body mass will be and the longer the culmen will be.

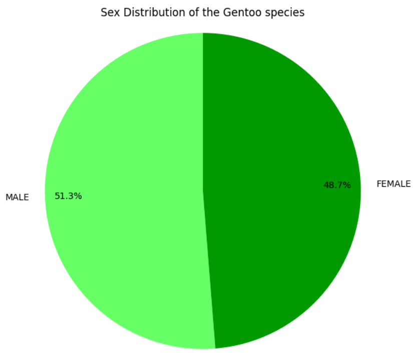
There is a negative correlation of -0.58 seen between flipper length and culmen depth which indicates that the deeper the culmen is of a given penguin, the shorter the flippers will be. And a negative correlation of -0.47 is seen between culmen depth and body mass, so a deeper culmen also indicates a smaller body mass.



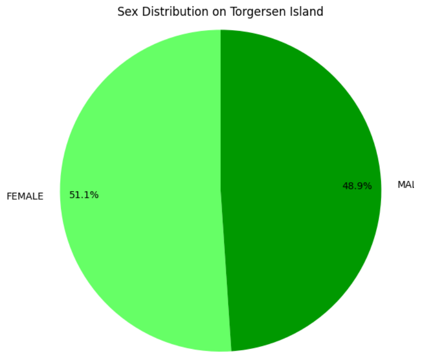
Finally the gender distribution was examined. The ratio of males to females across the 3 different species didn’t vary much, with the Adelie and Chinstrap species having a 50:50 ratio of males to females, and the Gentoo species having slightly more males (51.3%) than females (48.7%).

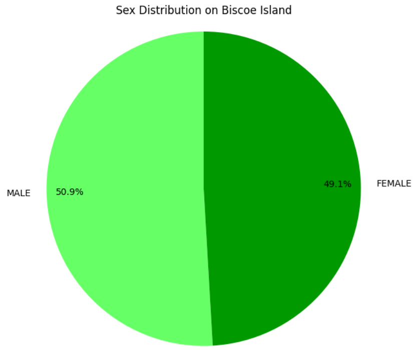


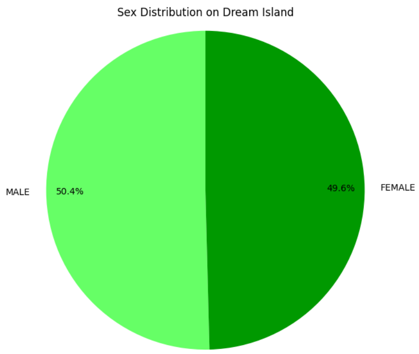




The gender ratios across the three different islands can be seen in the pie charts below.







THIS REPORT WAS WRITTEN BY : WILLIAM BIGWOOD