Hyperiondev

TASK

Exploratory Data Analysis on the Penguin Census Dataset

Visit our website

Introduction

This report is analysing a dataset containing various attributes related to penguins found on 3 islands in the Antarctic. The 3 islands that were surveyed were the Biscoe islands, Dream island, and Torgersen island.

After the penguins were captured, they had measurements taken and attributes recorded. These were the sex and species of the penguin, the culmen length and depth, in millimetres, the length of the flippers, also in millimetres, and the body weight of the penguin, in grams.

The culmen describes the ridge that goes along the maxilla (upper bill). Therefore, the length of the culmen is the distance from the base, to the tip of the maxilla. The depth of the culmen is defined as the distance between the highest point of the culmen and the lowest point of the mandible (lower bill).

DATA CLEANING

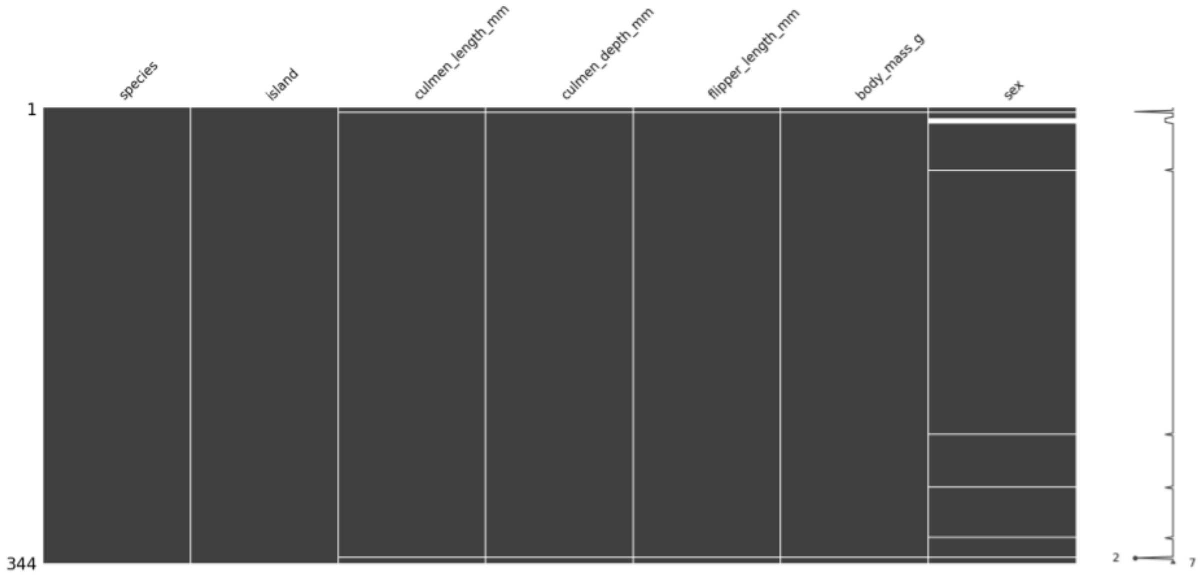

Figure 1: A plot showing all of the empty cells within the dataset

Initially, a matrix was generated to show all the empty cells within the entire dataset. This plot is shown in figure 1. The dataset starts with the vast majority of the cells filled. However, the sex column contains the most empty cells. Another point to note is that there are two rows with almost all of the attributes missing. After looking at all the unique values for each column, it was found that some of the values within the sex column had a value of ".". This is probably to indicate that the sex of the penguin could not be determined. All other values within the dataset were valid and useable.

MISSING DATA

The two rows that were almost completely empty may have been due to the fact that the scientists recording the data lost the penguin before being able to record anything of value apart from the species and the island that it was present on. This indicates that the missing values were not random. The rows missing the sex values could have been due to a similar issue.

When dealing with these inconsistencies, the approach taken was to drop the rows entirely. This was done because it would not affect the statistical outcomes significantly due to the ratio of empty cells being very small. Dropping all the rows that contain empty cells or that have a sex value that is neither male nor female succeeded at cleaning the dataset. The resulting set contained no empty cells and only valid data points.

DATA STORIES AND VISUALISATIONS

The analysis started by understanding the spread of different species across the three islands that were taken into account.

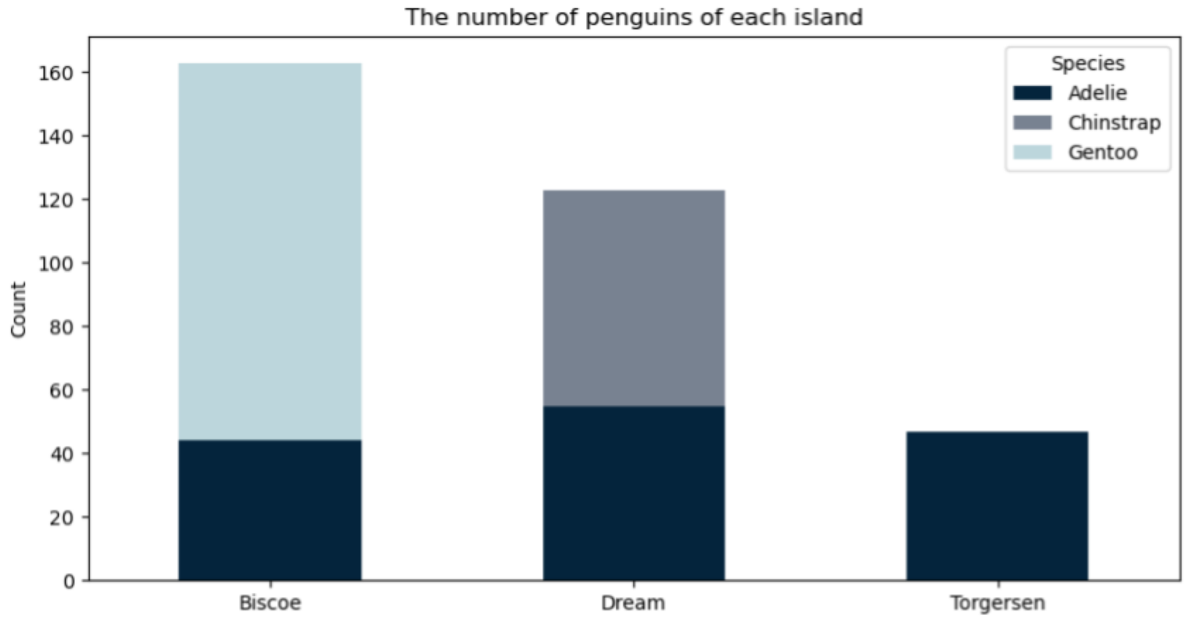

Figure 2: A bar plot showing the penguin population on the Biscoe, Dream, and Torgersen islands.

The outcome of this is shown in figure 2. Adelie penguins seem to be distributed relatively evenly between all three islands. Chinstraps were confined to Dream island at a similar ratio to the Adelie population on the same island. The Gentoo penguins were contained on Biscoe island, however, they made up a much higher population ratio compared to the Adelie also on Biscoe island. Torgersen only contains Adelie penguins.

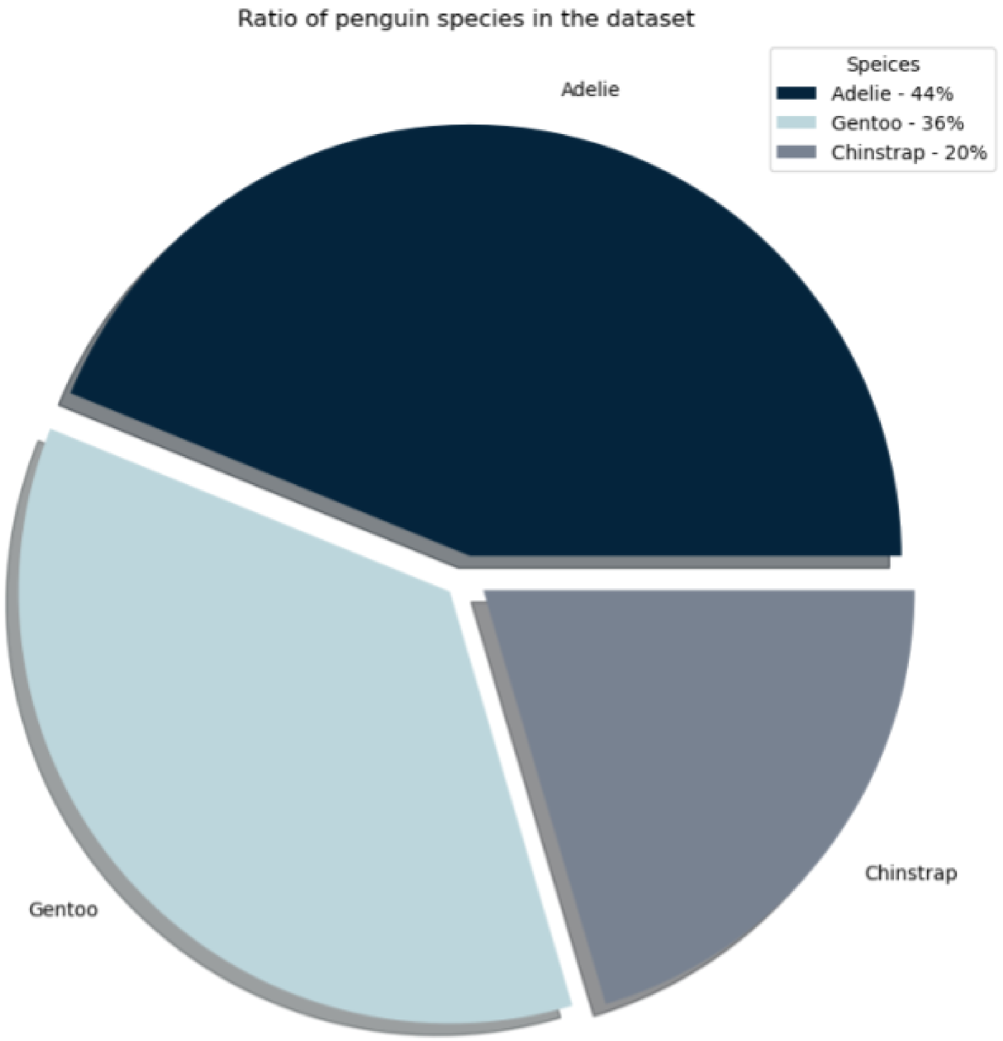

Figure 3: A pie chart showing the ratio of each penguin species within the dataset

Though the stacked bar chart shows the populations on each island, it does not do a satisfactory job of showing which species are the most dominant as a whole. This was fixed by plotting a pie chart, shown in figure 3. Adelies were the most common penguin to be recorded with a ratio of 44% of the entire dataset. This was expected as they were found on all three islands., Gentoo penguins were the second most common species with a ratio of 36%. Chinstraps were the least common with a ratio of 20%. This was also expected as the Chinstrap population was similar to the Adelie population specifically on Dream island and much smaller than the Gentoo population on Biscoe island.

The data exploration now leads to comparing different attributes of the penguins in relation to their species. It is anticipated that there will be significant differences in these attributes.

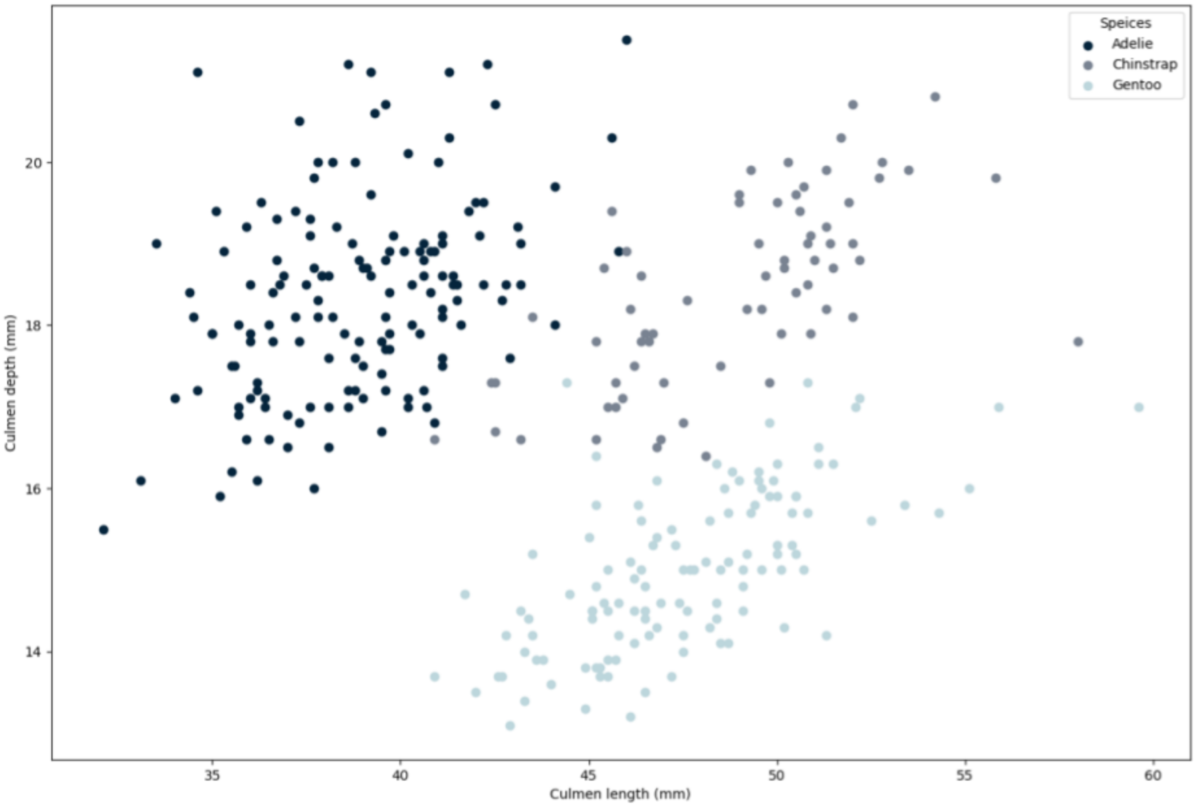

Figure 4: A scatter plot comparing the relation between the culmen length and depth, separated by species

Figure 4 is comparing how the culmen length correlates to the culmen depth. The points were also separated by species. Even though each species sees a separate cluster of points, they all show a positive correlation. This means that generally, the longer the culmen, the deeper the culmen, within the species. Due to the differentiation by species, more information can be extracted. For example, the Gentoo has similar culmen lengths to the Chinstraps. Although this is the case, the Gentoo have significantly shallower culmens. A similar comparison can be made for the Adelie and the Chinstraps. Both have similar culmen depths, but the Adelie have much shorter culmens in general. Therefore, the Chinstraps and the deepest and longest culmens overall.

After finding how the species culmen measurements compare between eachother, analysing how the sex affects the measurements was the next logical step. To distingush whether this difference was species dependent, the species had to be plotted individually.

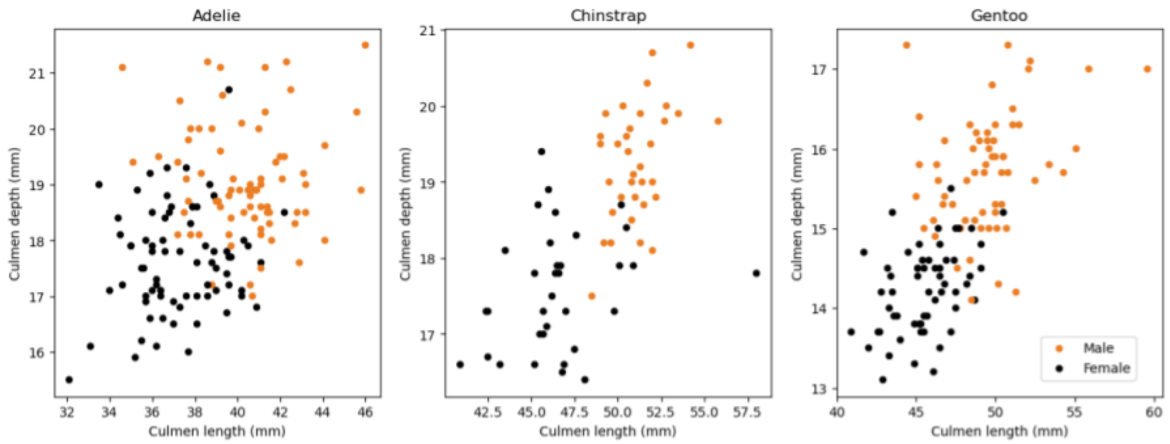

Figure 5: Scatter plots showing how the sex of the penguin affects the culmen measurements for each species.

Figure 5 shows this expansion on the culmen analysis. The positive correlation for each species is much more noticeable after the separation. The main trand in this visualisation is that the male penguins tend to have longer and deeper culmens compared to the females. This trend holds for all three speices. There is the least overlap between males and females in the Chinstrap population. Though this could be due to the fact that the Chinstrap population is also the smallest.

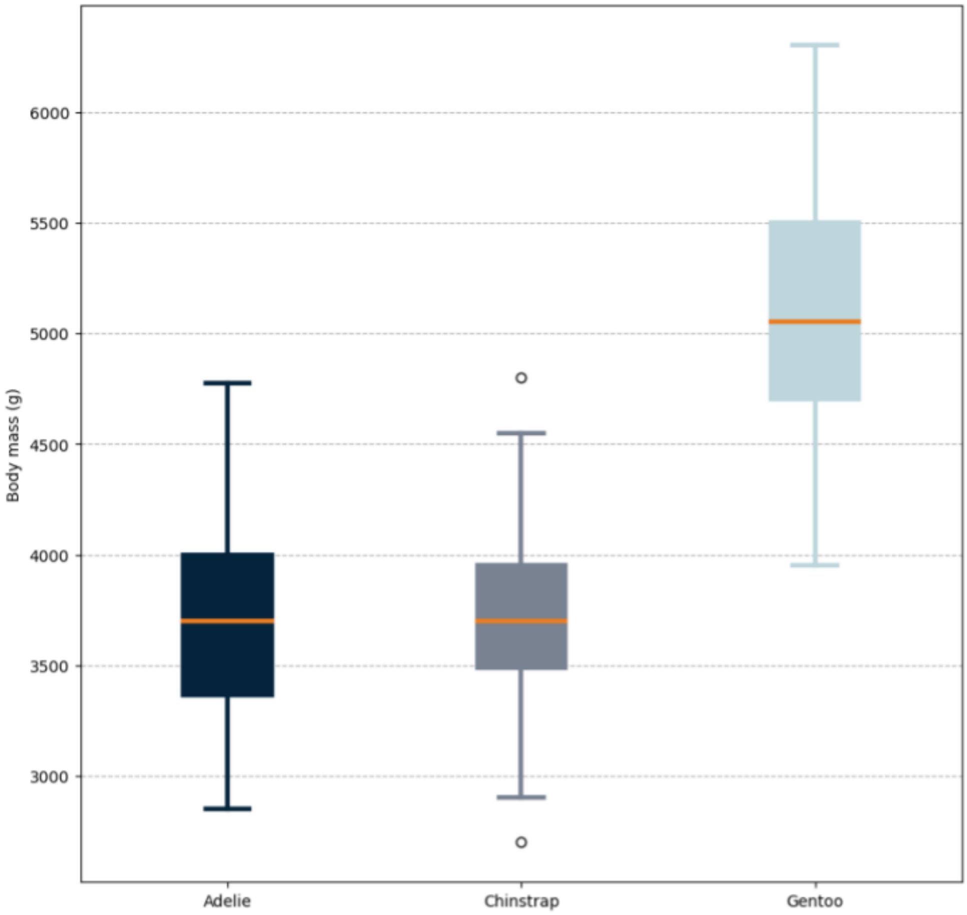

Figure 6: Box plots showing the body mass distribution of the Adelie, Chinstrap, and Gentoo penguins.

After collecting the information about the the culmen measurements, the other attributes were focused on. Figure 6 shows the distribution characteristics for the body mass of each penguin species. One of the obvious trends is that Gentoo penguins are heavier than the Adelie and Chinstraps. Comparing the spread, Gentoo penguins have a much wider range. Adelie and Chinstrap's body mass average are very much alike, though the Chinstraps have a narrower distribution. However, this narrowness means that outliers are present, where the other species dont show any.

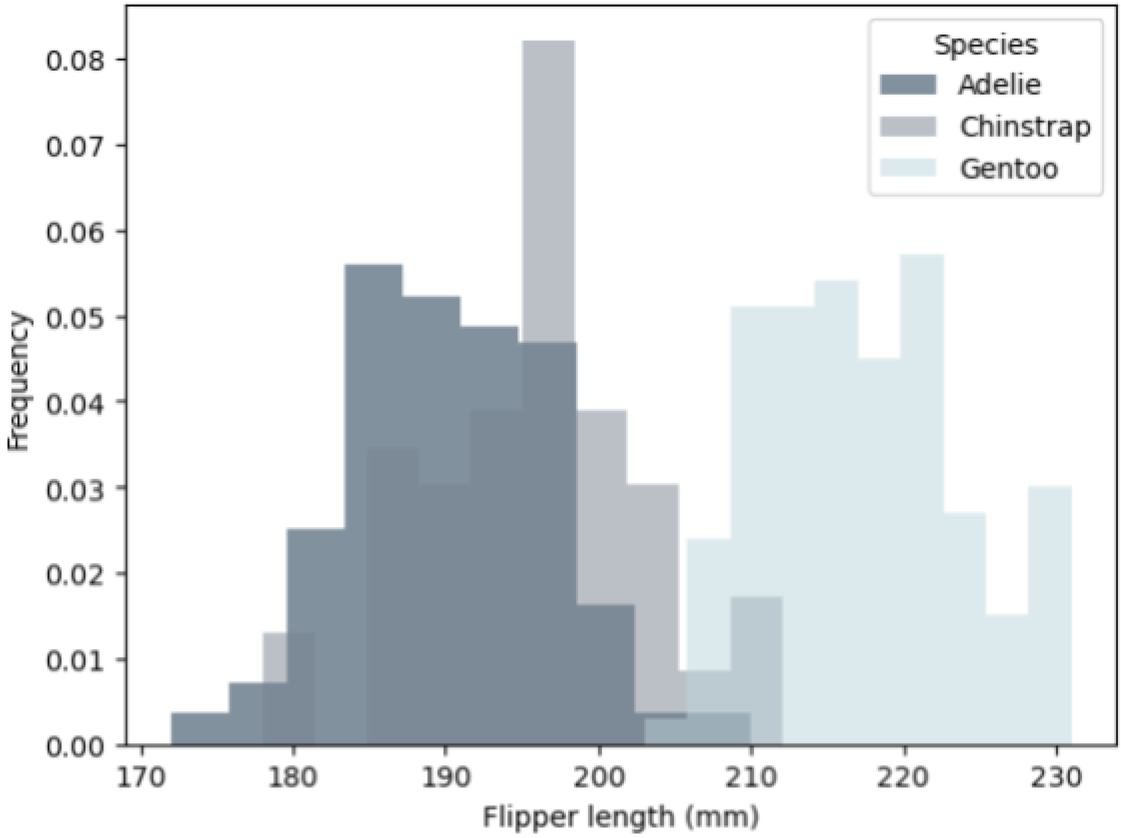

Figure 7: Overlapping histogram showing the distribution of flipper lengths between Adelie, Chinstrap, and Gentoo penguins.

The next part of the analysis was to identify the differences between flipper lengths according to species. Figure 7 was made to accomplish this. The Gentoo penguins are significantly longer than both other species. This makes sense since the Gentoo is also the heaviest, and their culmens are some of the longest. All attributes pointing to Gentoos being the largest of the pegnuins. Chinstraps have the next longest flippers, but they are still similar to the Adelie flippers. However, the number of Chinstrap penguins with the same flipper length is much higher than that of the Adelie or the Gentoo.

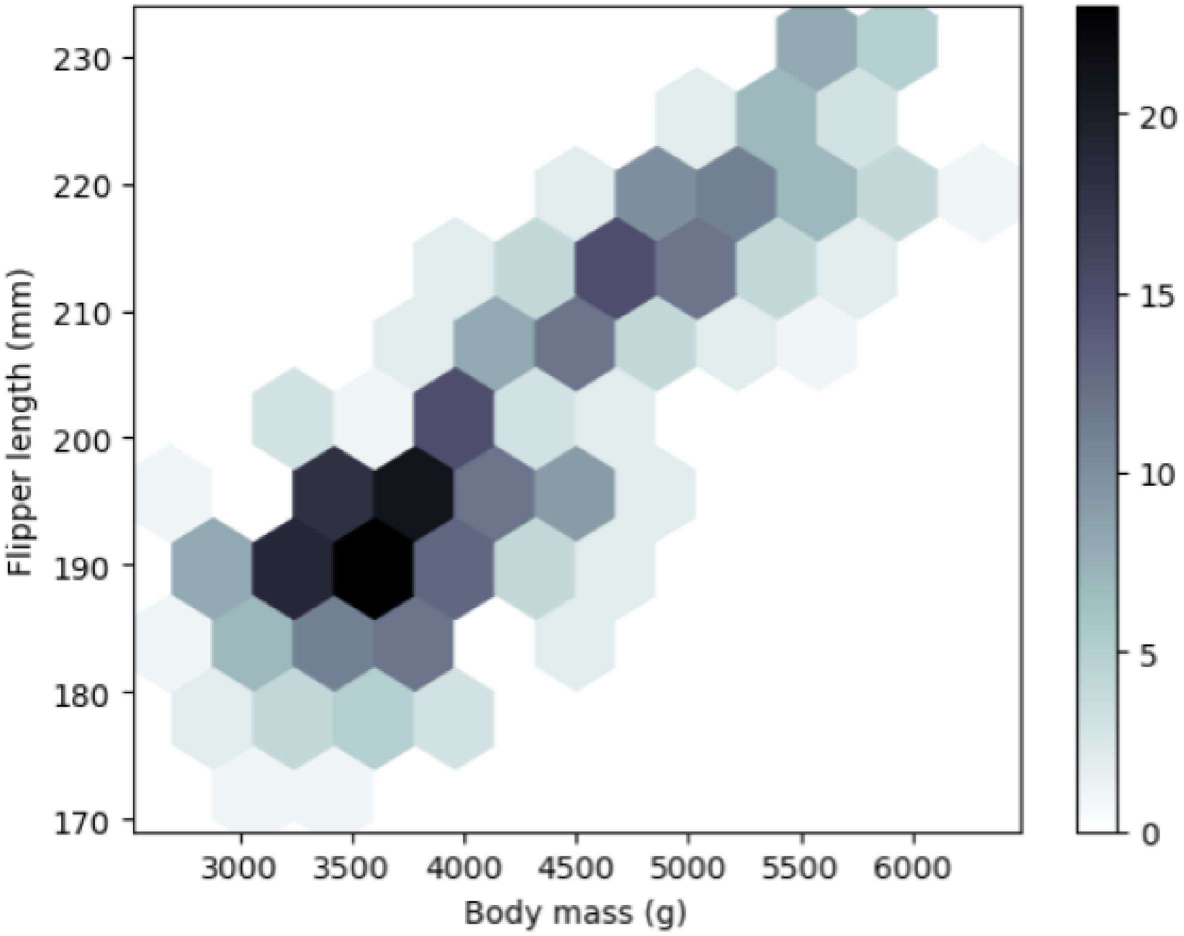

Figure 8: A hex 2D histogram showing the relation between flipper length and body mass.

To test the hypothesis that longer flippers were expected for heavier penguins, this relation was plotted across the entire dataset, figure 8 was generated. This shows that there is a positive correlation between these two variables. Therefore, with heavier penguins, the longer the flippers. The plot also shows that there are two areas of high density (dark portions). The higher density area with the lower mass and shorter flippers is due tio the fact that the Adelie and Chinstrap penguins have similar measurements in these attributes. The other high density space is where most of the Gentoo penguins lie. This is way that area is not as dense.

To summarise, the Adelie penguins are the most common, existing on all three islands. They also are similar in body structure to the Chinstrap in the sense that they haVe a similar mass and flipper length. I am going to assume this means that they are generally the same height aswell. However, they have different bill profiles, with the Adelie having shorter culmens. The Gentoo penguins are the heaviest and have the longest flippers. This leads me to believe that Gentoo penguins are probably larger and taller than the other species. Their culmens are some of the longest, but also some of the shallowest. The Chinstraps share culmen characteristics with both other species. For example, their culmens are the same length as the Gentoo, but the same depth of the Adelie.

THIS REPORT WAS WRITTEN BY: Conner Grice