Hyperiondev

TASK

Exploratory Data Analysis on the Penguin Data Set

Visit our website

Introduction

This data set contains information from 3 studies that recorded nesting observations, penguin size data, and isotope measurements from blood samples for adult Adélie, Chinstrap, and Gentoo penguins.

DATA CLEANING

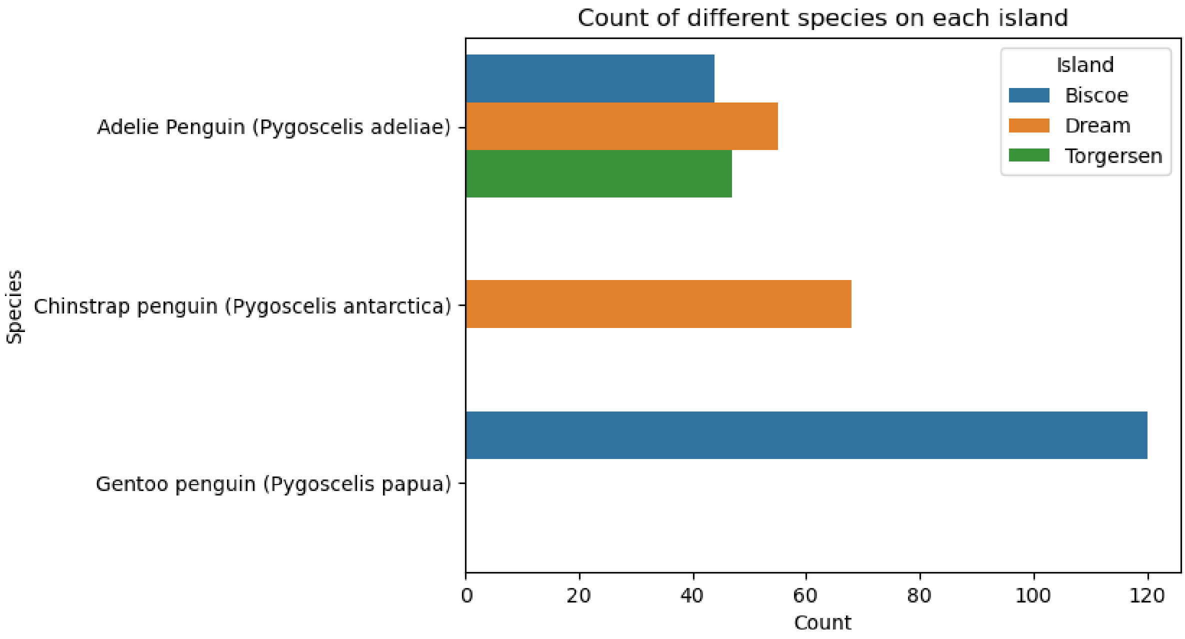
For our analysis, we have decided to drop the isotope measurements from the data set as we will not be analysing this information. At the same time, we removed any duplicate rows from the data set. We then made sure that the remaining columns were the correct data type so that we could properly analyse the information. The only column that needed attention was the Date Egg column which was initially just a string so we have parsed it into a DateTime format.

MISSING DATA

We then looked into finding any rows with missing data, the comments section had lots of missing entries but this is as expected as we are not expecting every entry to need a comment. The next column that had missing data was the Sex column and as there was no way we could try to infer the value for the missing entries we decided to remove these rows from the data set. When we removed these rows there were no more rows with missing data so we can now start our analysis.

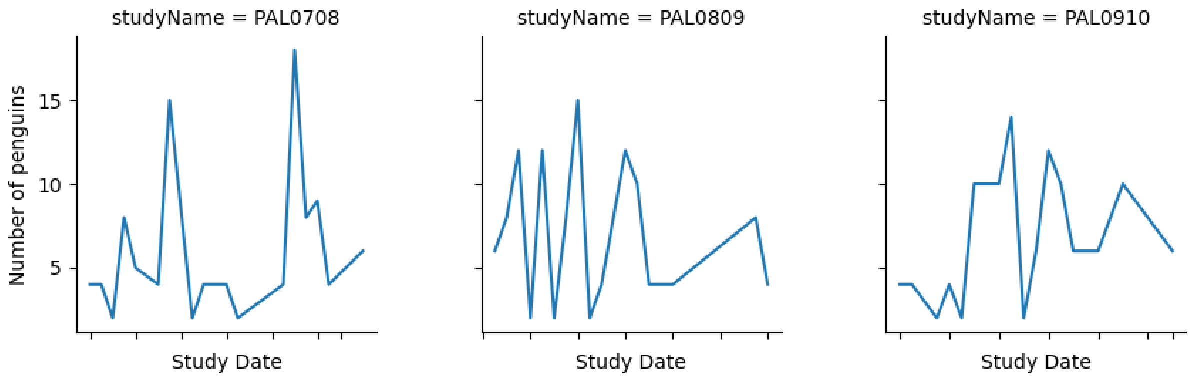
DATA STORIES AND VISUALISATIONS

The first thing we want to analyse is to find how many different penguins of each species were found and where were they found. To investigate this we made the following graph.



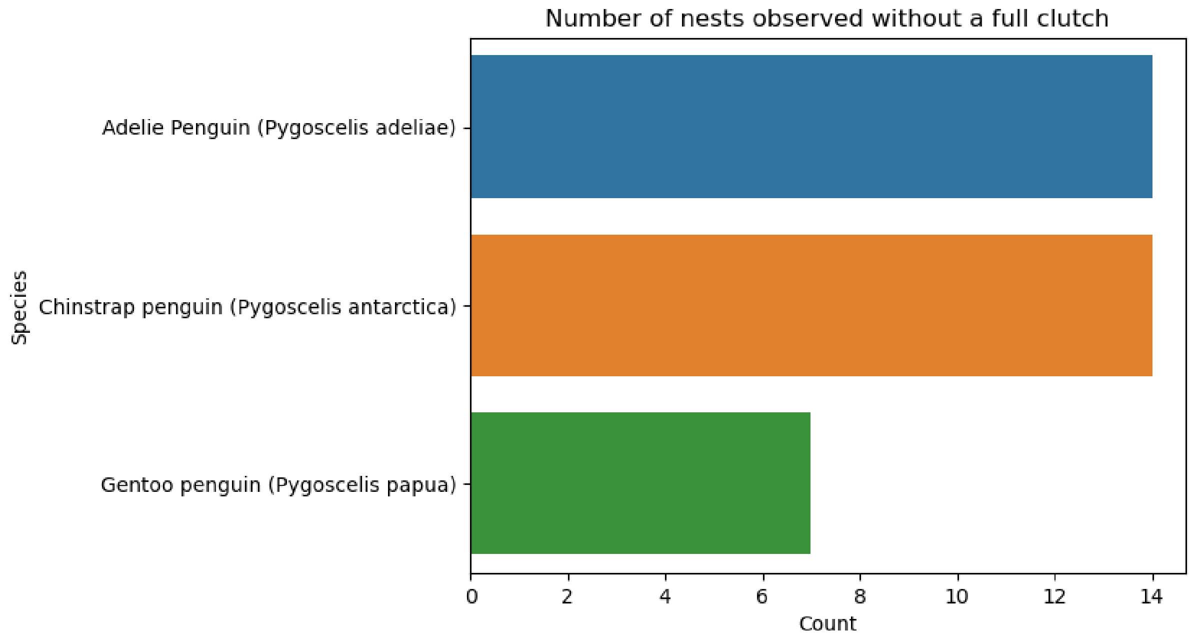
Here we can see that the island with the most penguins was Biscoe, and the most common penguin species was the Adelie Penguin. The Adelie Penguin is also the only penguin that was found on more than one island.

Next, we want to see how many penguins were found each day for each study. To do this we have created the following graph.



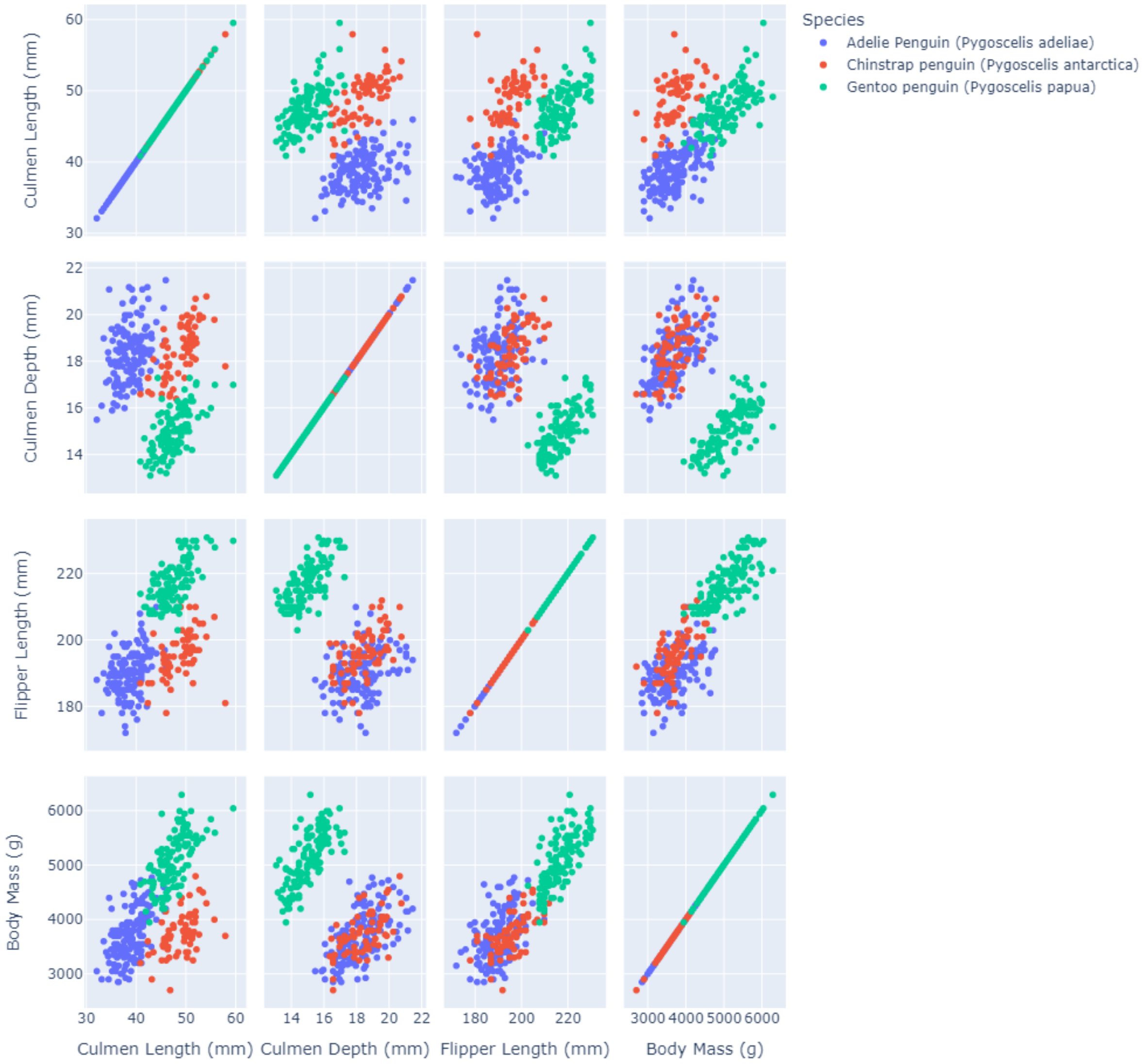
Here we can see that all 3 studies went through peaks and troughs with the number of penguins they observed on each day of the studies. The first study ‘PAL0708’ has the biggest spikes where had two days where they observed over 15 penguins in a day and then much less in the rest of the study whereas the other two had smaller spikes but consistently more observed each day.

Now we want to see how many times a nest was observed without the full number of eggs being there and to see how it varies with each species.



Here we can see that the Adelie and Chinstrap had the most number of nests without a full clutch of eggs. It would be expected to see Adelie have the highest as it was the most commonly observed penguin, but, surprisingly, the Chinstrap has the same amount when there are considerably fewer of them observed. There may be some external cause that is impacting the Chinstrap Penguins.

Finally, we want to explore how Culmen length, Culmen diameter, Flipper length and Body mass vary between the three different species.



From this graph, we can see that the Gentoo penguin is the larger species with it having a greater body mass, flipper length and culmen length but it does have a smaller Culmen depth compared to the other two species.

THIS REPORT WAS WRITTEN BY : Tom Anderson