

**TASK**

**Exploratory Data Analysis on the Penguin Data Set**

[Graphical user interface, text

Description automatically generated](https://www.hyperiondev.com/)

# Introduction

This report presents an analysis of the "penguin\_lter.csv" dataset. The dataset contains a wide variety of information regarding the measurements and isotopic signatures of various species of adult male and female penguins residing on islands near Palmer station along the Palmer Archipelago, Antarctica. The data has been imported and organised into a structured format, referred to as the "penguins\_df" dataframe.

The dataset contains information on species of penguins, including:

* Species and sex: The three species of penguin studied: Adelie (Pygoscelis adeliae), Chinstrap (Pygoscelis antarctica), and Gentoo (Pygoscelis papua). The sex of the penguin was determined by blood samples.
* Region and island: The geographical location of the penguin colonies. The individuals studied resided in the region of Anvers specifically on three different islands: Torgersen, Dream, and Biscoe.
* Nest information: Includes information on clutch completion (if the study nest observed with a full clutch, i.e., 2 eggs) and date egg (the date the study nest observed with 1.
* Size and weight measurements: physical measurements of the penguins, such as the length and depth of the culmen (the upper ridge of the bill), the length of the flippers, and the body mass. These measurements are important for understanding the physical characteristics of the different species of penguins and how they may vary.
* Isotopic information: the isotopic signatures of the penguins, such as the levels of carbon and nitrogen isotopes in their blood. These isotopic signatures can provide information about the penguins' diet and foraging habits, as different prey species have different isotopic signatures. This information is useful in understanding the ecology of the penguin colonies and how they interact with their environment.

# Data Cleaning

In the process of preparing the "penguins.csv" dataset for analysis, several data cleaning steps were applied. Columns deemed irrelevant or unnecessary were removed from the data frame. These columns were 'Region', 'Stage', 'Individual ID', 'studyName', 'Sample Number', and 'Comments'.

On inspection of the data types of the columns, it was seen that ‘Date Egg’ column was classed as an object. The column was therefore transformed into a datetime format. In addition, a new column was created called ‘Year Egg’ stating the year only.

# Missing Data

On inspection of the columns, it was seen that columns 'Culmen Length (mm)', 'Culmen Depth (mm)', 'Flipper Length (mm)', 'Body Mass (g)', ‘Sex’, ‘Delta 15 N (o/oo)’, and ‘Delta 13 C (o/oo)’ contained missing data. In addition, the ‘Sex’ column contained a ‘.’ which was treated as missing data and so, converted to a null value. The missing values were imputed using a most frequent strategy.

# Data Analysis and Visualisation

Figure 1 shows count plots for various species of penguins, separated by sex and the island in which they resided. The figure 1a shows that Adelie penguins are the most common in the dataset and Chinstrap are the least common. There are also slightly more male Adelie and Gentoo penguins and an equal number of male and females Gentoo penguins. Figure 1b shows Adelie penguins are located on all three islands. Chinstrap penguins are only documented on Biscoe Island and Gentoo penguins are only located on Dream Island.

Chart, bar chart

Description automatically generated

**Figure 1.** Cluster Count plots for Species versus Sex and Island

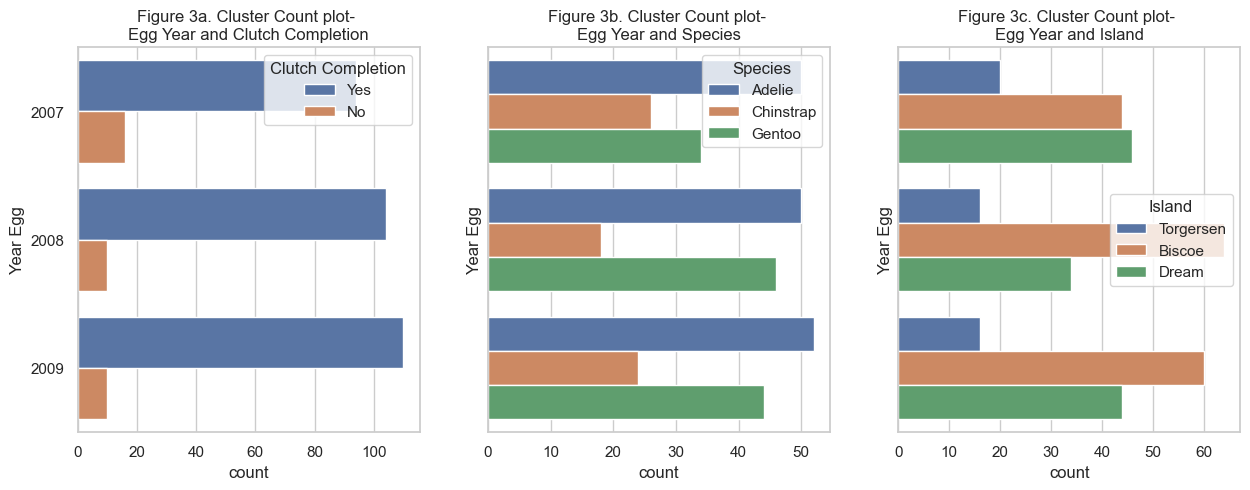
Figure 2 shows the number of eggs observed for the three years. All eggs were observed in November/early December suggesting that this is the time of the year when the eggs are laid. The most eggs observed in one day was in 2007.

Figure 3 shows count plots for the year in which an egg was observed. Figure 3a shows the plot clustered by whether a complete clutch was observed. The plot shows that most of the nests observed are complete. It shows that each subsequent year saw an increase in a complete clutch. Figure 3b shows the plot clustered by species. For all three years, Adelie penguins laid the most eggs and Chinstrap, the least. Figure 3c shows the plot clustered by the Islands in which the penguins resided. Torgersen island had the lowest number of eggs laid. In 2007, Dream Island had the highest number of eggs laid and in 2008 and 2009, Biscoe Island had the highest number of eggs laid.

Chart, line chart

Description automatically generated

**Figure 2.** Line plots for the number of eggs observed for each year



**Figure 3.** Cluster Count plots for Egg Years versus Clutch Completion, Species, and Island

Figure 4 shows Violin plots for the size and weight measurements- Culmen Length, Culmen Depth, Flipper Length, and Body Mass- separated by species. The violin plot operates as a combination of a boxplot and a kernel density estimate. The graphs show the similarities and differences of the three species, based on their physical properties. Figure 4a shows that Adelie penguins have the smallest culmen length and the distribution on culmen lengths. Chinstrap and Gentoo penguins are similar in length with Chinstrap penguins having the highest median length. Chinstrap and Gentoo penguins are also bimodal in shape. As shown in figure 5, this is likely due to the sexual dimorphism of the penguin species. Using culmen length as an example, figure 5 shows that certain physical properties of the same species can differ, depending on sex, and as such, distinguishing features could be used to differentiate between the sexes. Figure 4b shows the violin plot for culmen depth. Gentoo penguins have the lowest culmen depth and Chinstrap penguins have the highest median depth. Adelie penguins have the largest max depth.

Shape

Description automatically generated

**Figure 4.** Violin plots for Species versus Size and Weight measurements

Chart, radar chart

Description automatically generated

**Figure 5.** Violin plots for Species sex versus Culmen Length

Figures 4c and 4d show that Gentoo Penguins have the longest flipper length and body mass, while Adelie penguins have the shortest flipper length and both Adelie and Chinstrap penguins are similar in weight.

Figure 6 shows a scatter matrix of the size and weight measurements. This plot is effective as a means of differentiating different species based on their physical properties. For example, figure 6g shows the grouping of Gentoo penguins are separated from the groupings of the other two species meaning that Gentoo penguins can be easily determined from those measurements. Similarly, plots such as figures 6e, 6i, and 6m have a high degree of specificity and may be used to accurately determine the species type. Similarly, a scatter matrix can be used to determine the sex of a specific species, based on the physical properties. Figure 7 shows scatter plots for the three species with physical properties showing high degrees of separation between the two sexes.

Scatter chart

Description automatically generated

**Figure 6.** Scatter Matrix for Size and Weight measurements

Chart, scatter chart

Description automatically generated

**Figure 7.** Scatter plots for Differentiation based on Sexual Dimorphism

Figure 8 shows violin plots of isotopic information based on species. Both plots are a measure of the isotopic composition of nitrogen and carbon in a sample. it is used to express the difference in the ratio of a heavy isotope (15N and 13C) and the reference isotope (14N and 12C). Delta 15N values can be used to study the trophic level of organisms, which is the position of an organism in a food chain. It is known that generally, the higher the trophic level of an organism, the higher the delta 15N value will be. Figure 8a shows that Chinstrap penguins have the highest median delta 15N and Gentoo have the lowest. Delta 13C values can be used to study the carbon cycling in ecosystems and to trace the origin of carbon in the food web, and it can be used to study the metabolic pathways of organisms, and also to study the water use efficiency of plants. Figure 8b shows that Chinstrap penguins have the highest median delta 13C and Gentoo and Adelie have the lowest with similar values.

Chart, radar chart

Description automatically generated

**Figure 8.** Histograms for Isotopic information

THIS REPORT WAS WRITTEN BY: Joshua Houlden

