

**TASK**

**Exploratory Data Analysis on the \*\*\*\*\*\*\*\*\*\*\*\* Dataset**

[](http://www.hyperiondev.com/portal/)

**Introduction**

Summary of the data set

The dataset provides various attributes of penguins, specifically focusing on three species: Adelie, Gentoo, and Chinstrap. The dataset includes numerical features such as body mass, flipper length, and culmen depth. It allows for various kinds of exploratory and statistical analysis aimed at understanding patterns and variations in these penguin species.

The data collected has been manipulated in a way that has helped us identify certain characteristics of each penguin species and also find correlations between certain traits.

**DATA CLEANING**

The data was cleaned in a way that makes it easier to plot accurate graphs and subsequently provide accurate data analysis. In this task I cleaned the data by making sure that numerical data was correctly identified as numerical by using the .to\_numeric( ) command. I also cleaned the data by dropping any duplicate rows. I did this by using the .drop\_duplicates( ) command.

**MISSING DATA**

There was missing data in the dataset. This was in the form of “NA” data points. These NA points were converted to NaN points and then dropped using the .dropna( ) command.

**DATA STORIES AND VISUALISATIONS**

1. Sampling Size by Species (Bar Plot):

**A graph of different species

Description automatically generated**

Findings:

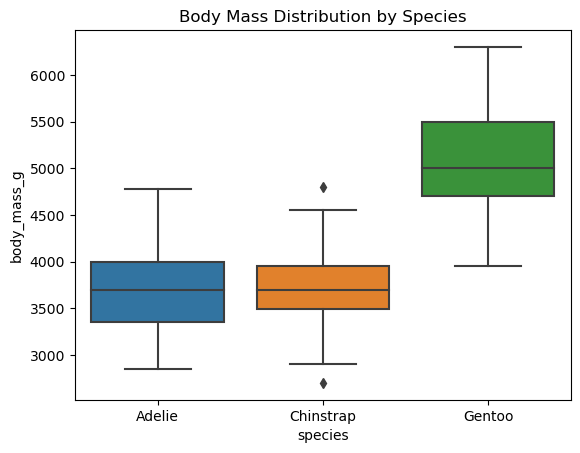
Adelie penguins are the most represented in the dataset, followed by Gentoo and then Chinstrap.

The sampling size of the species doesn't necessarily reflect the actual abundance of these species in the wild.

Insights:

This might indicate a potential sampling bias. Understanding the context in which the data was collected is crucial. For instance, if the study aimed to get an even distribution across species, then this result is indicative of a deviation from that goal. On the other hand, if the dataset is simply a collection from a certain location, then it might reflect the local population distribution.

1. Distribution of Body Mass by Species (Box Plot):



Findings:

Gentoo penguins are generally the heaviest species in the dataset.

The median body mass for Adelie and Chinstrap penguins is relatively close.

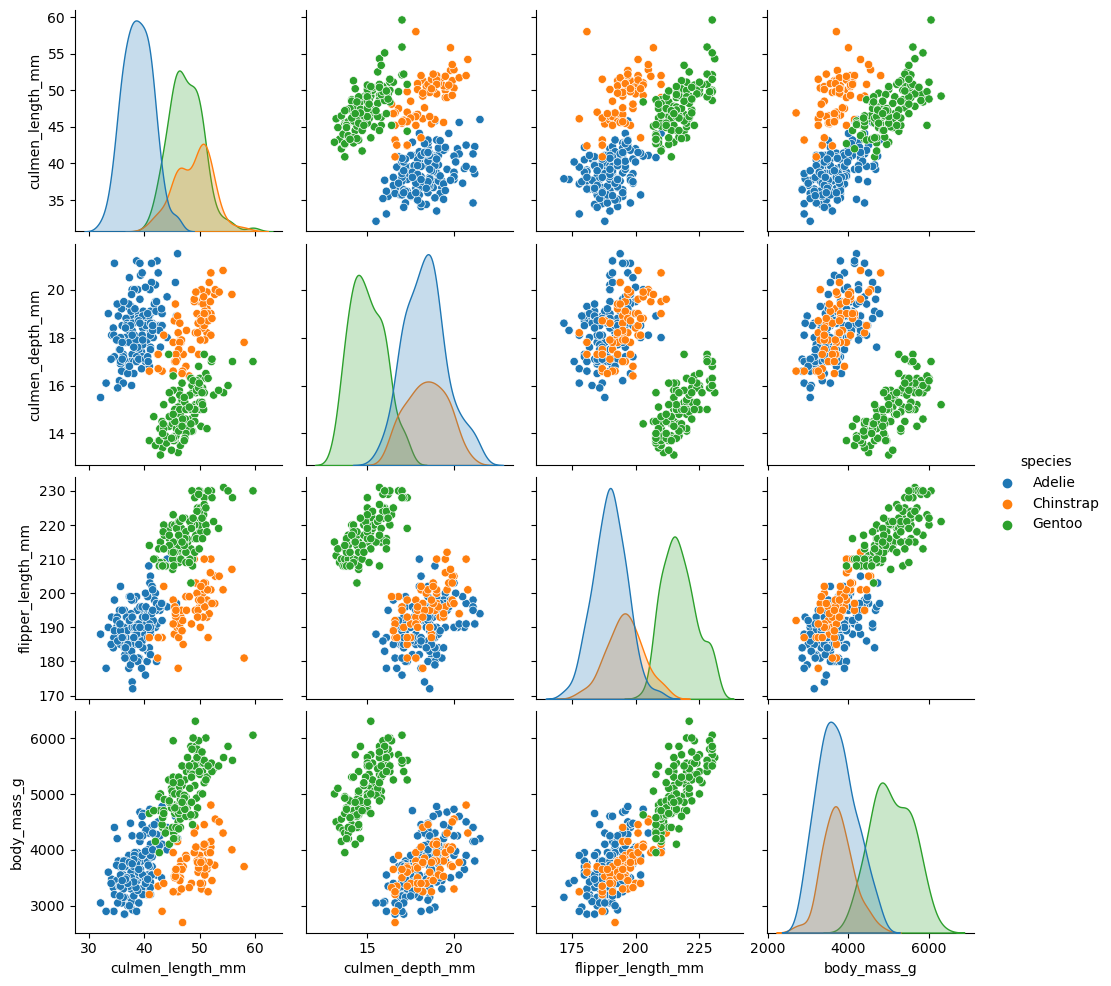
Chinstrap penguins have outliers in their body mass distribution.

Insights:

The higher body mass in Gentoo penguins could suggest they have a different diet, habitat, or evolutionary adaptation compared to Adelie and Chinstrap penguins.

The presence of outliers in the Chinstrap penguins could point to potential anomalies, errors in data collection, or rare instances of unusually heavy or light individuals. Further investigation is needed.

1. Species Relationship (Pair Plot):



Findings:

Direct correlation between body mass and flipper length.

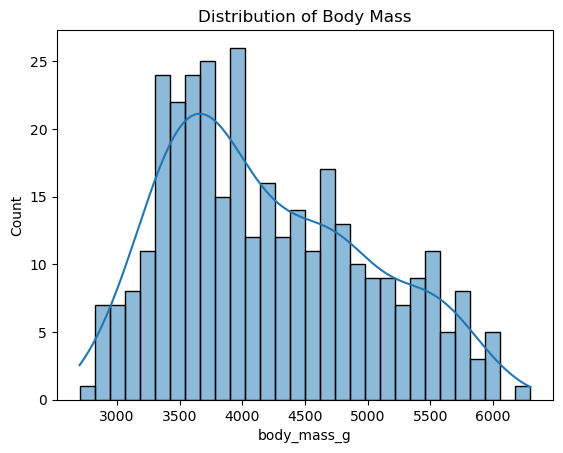
Inverse correlation between flipper length and culmen depth.

Insights:

The correlation between body mass and flipper length might suggest evolutionary adaptations. For instance, heavier penguins might need longer flippers for efficient swimming.

The inverse relationship between flipper length and culmen depth might indicate a trade-off in the evolution or development of these features.

1. Distribution of Body Mass Across All Penguins (Histogram):



Findings:

The majority of penguins in the dataset have a body mass ranging from 3400 - 4000 grams.

Outliers are observed at 6200 grams and 2300 grams.

Insights:

This distribution provides a central tendency of body mass for the general penguin population in the dataset.

The outliers might indicate unique individual circumstances, errors in data collection, or genuinely rare instances of exceptionally heavy or light penguins.

**THIS REPORT WAS WRITTEN BY: Rameez Bhayed.**

