

**TASK**

**Exploratory Data Analysis on the Penguin Dataset**

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

**Introduction**

This is a dataset comprising various measurements of three different penguin species, namely *Adelie, Gentoo, and Chinstrap*. In the given dataset, the culmen refers to the “the upper ridge of a bird’s beak” of the penguin. In the raw data, bill dimensions are recorded as "culmen length" and "culmen depth".

**Data Methodology:**

As a first step we import necessary libraries, such as

1. Pandas for exploration,
2. Seaborn, Matplotlib for data visualization.

Next we read the dataset for further analysis.

**DATA CLEANING**

Dataset contains NaN (Not a Number) values in few columns which is cleaned for our analysis. We can observe that

1. Each float category column has 2 missing null observation
2. Sex column has 10 missing null values.
3. Bad data in sex column which needs to be addressed.

**MISSING DATA**

Handling Null values:

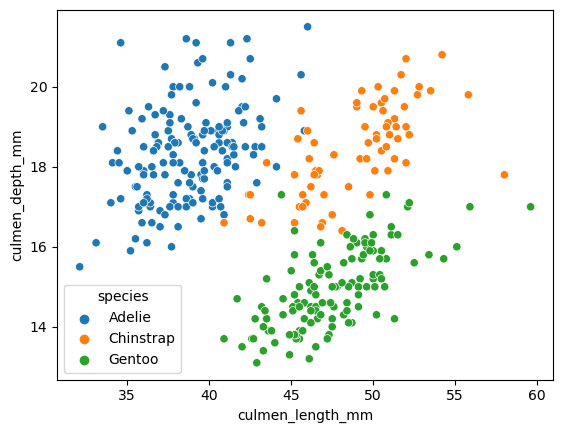
Certain columns contains null values.

1. **body\_mass\_g:** float category column has 2 missing values, except for species and island columns all other fields have null values and hence two fields are dropped.
2. **culmen\_length\_mm:** As the null values in body\_mass\_g column are dropped, culmen length column has been cleaned without null values.
3. **culmen\_depth\_mm:** As the null values in body\_mass\_g column are dropped, culmen depth column has been cleaned without null values.
4. **flipper\_length\_mm:** As the null values in body\_mass\_g column are dropped, flipper length column has been cleaned without null.
5. **Sex:** field has two different type of missing data, 9 columns with null values and one column with ‘.’ in its field.

All null fields are replaced with mode value of that column.

**DATA STORIES AND VISUALISATIONS**

The scatter plot of culmen lengths and depths by penguin species.



From the above chart, there are more Adelie species with culmen depth penguins

When points are graphed on a scatterplot, it is possible to find a line of best fit—a straight line that best represents the data on a scatterplot.

**Adelie:**

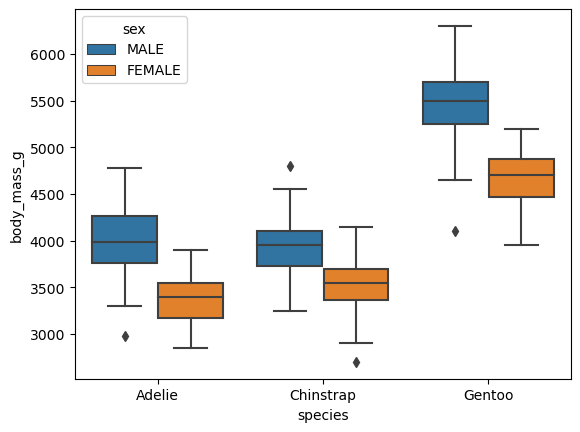
* + Adelie demonstrates weak positive correlation between the culmen depth and the length.
  + There are a bunch of data elements demonstrating to be anomalies. It doesn't fit the pattern of the other points and we wouldn’t be able to draw the line of best fit.
  + There is area of concentration between 35-45 length and the depth scattered between 16-20 mm.

**Chinstrap:**

* Chinstap demonstrates relatively a positive correlation between the culmen depth and the length.
* There are few data elements demonstrating to be anomalies. It doesn't fit the pattern of the other points and we wouldn’t be able to draw the line of best fit.
* There is area of concentration between 50-55 length and the depth scattered between 16-20 mm.

**Gentoo:**

* Chinstap demonstrates positive correlation between the culmen depth and the length.
* There are few data elements demonstrating to be anomalies. It doesn't fit the pattern of the other points and we wouldn’t be able to draw the line of best fit.
* There is area of concentration between 40-50 length and the depth scattered between 14-16 mm.

**Boxplot:**  


From the above boxplot, there are more Gentoo Male and the female species with maximum body mass compared to the Adelie and Chinstrap species.

When points are graphed on a boxplot, it would be evident to identify the Max, Min and the Median visually across the gender and the body mass. Also we can identify if there are any outliers for the given dataset.

**Adelie:**

* + Male Species: Adelie has a maximum at ~4750 and minimum at ~3250. Median lies @ 4000 and there is one outlier at 3K
  + Female Species: Adelie has a maximum at ~4k and minimum at ~2750. Median lies @ 3500 and there are no specific outliers plotted.

**Chinstrap:**

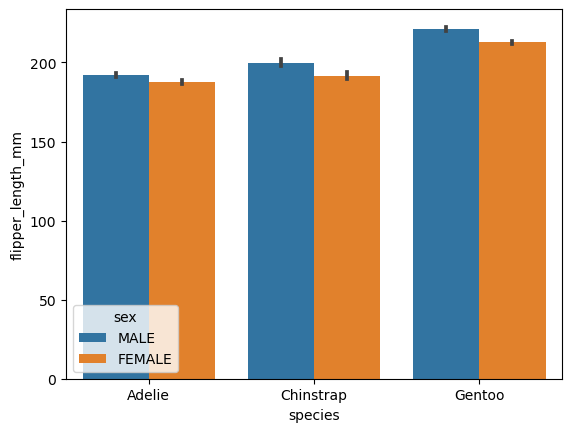
* + Male Species: Chinstrap has a maximum at ~4500 and minimum at ~3250. Median lies @ 4000 and there is one outlier at ~4750
  + Female Species: Chinstrap has a maximum at ~4250 and minimum at ~3000. Median lies @ 3650 and there is one outlier at 2K

**Gentoo:**

* + Male Species: Gentoo has a maximum at ~7k and minimum at ~3250. Median lies @ 5500 and there is one outlier at 4K
  + Female Species: Gentoo has a maximum at ~5250 and minimum at ~4k. Median lies @ 4750 and there are no specific outliers plotted.

Bar plot:

The flipper lengths of the species by sex



As a first insight from the Bar Plot we are able to infer that

There aren’t material difference between the Male vs Female across the species referenced

The flipper lengths of their male counterparts are greater than their female across the species marginally by 1-2 mm.

**Adelie**:

Adelie has a maximum at ~190 and the female’s flipper length is also in the similar range. Of the three species they are at the lower end of the flipper length.

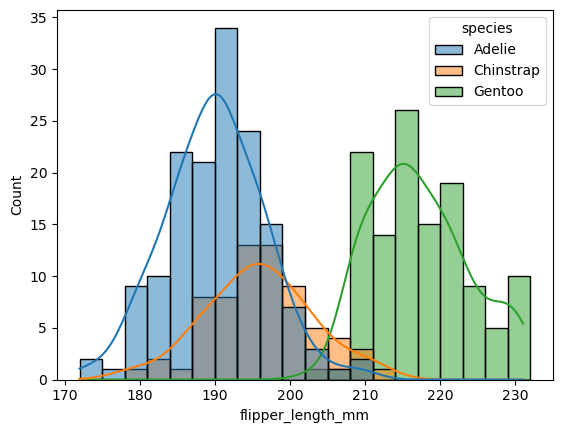
**Chinstrap**:

Chinstrap species has a maximum at ~200mm and the female’s flipper length is also in the similar range and slightly lesser than their male counterparts. Of the three species they are at the mid end range of their flipper length.

**Gentoo**:

Gentoo species has a maximum at ~230 and the female’s flipper length is also in the similar range and slightly higher than ~200mm. Of the three species they are at the higher range of their flipper length.

**Histogram:**



As a first insight from the Histogram we are able to infer that

* + There are two peaks spanning between Adelie and Gentoo species.
  + There are no major outliers across the species data

Distribution line Fit:

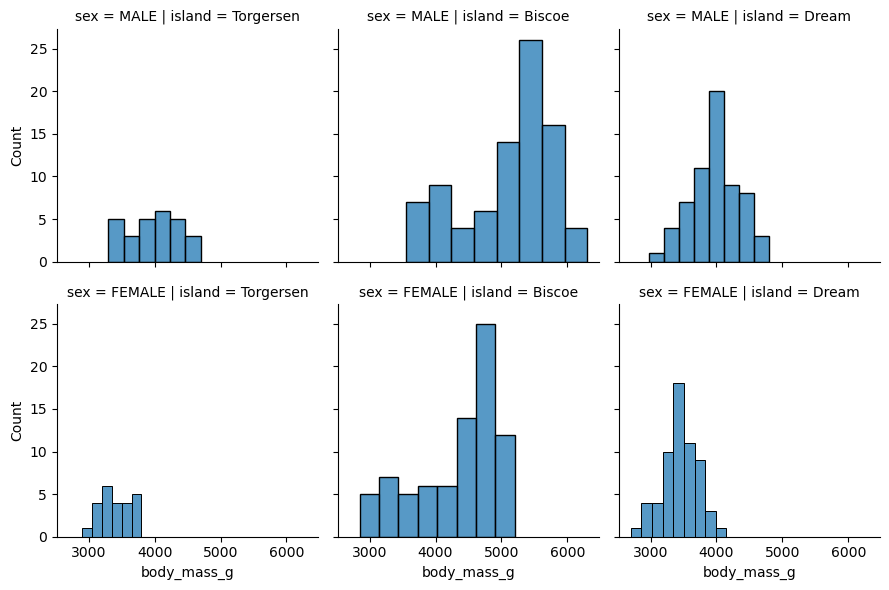
* **Adelie** : Relatively good fit
* **Chinstrap**: Right Skewed and poor fit
* **Gentoo** : Neither right / left skewed and not a good fit

**Adelie**: There are about 34 species which spans within 190 mm length and follows almost a near normal distribution for the sample data set.

**Chinstrap**: Chinstrap species shares the maximum length between 195-198 for about 13 species.

**Gentoo**: There are about 26 species which spans within 215 mm length and follows neither right / left skewed and not a good fit.

**Seaborn FacetGrid:**



As a first insight from the above plot we are able to infer that

1. There is wider distribution of male species in Biscoe Island and relatively similar female population in the same island.

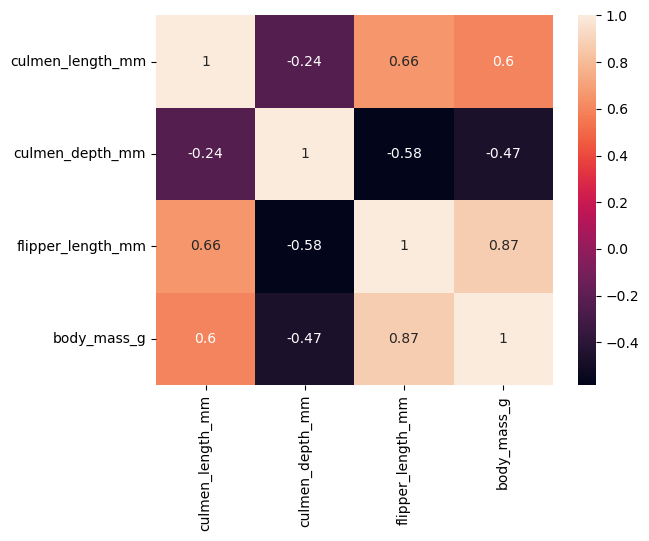
2. Torgersen Island has slightly lesser population than the other two islands under

Comparison for the sample population at study.

3. In Biscoe island both the male and female population seems to be right skewed and doesn’t have a normal plot

4. In Dream Island the Male species have normal fitting curve and the female species almost have a normal fitting curve.

**Heatmap:**



As a first insight from the heatmap we can infer that

1. There is deeper +ve trend between flipper length and body mass. Additionally, there is positive trend (0.66) for the flipper length and culmen length and similar pattern is observed for body mass and culmen length. (0.6)

2. There is negative trend for (-0.58) for culmen depth and flipper length

**THIS REPORT WAS WRITTEN BY: KALAISELVI RAJASEKAR**

