

**TASK**

**Exploratory Data Analysis on the Penguin Dataset**

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

**Introduction**

Penguin is a very useful dataset especially when learning Data Science in terms of its simplicity, limited number of columns and also the amount of different visualizations it provides.

The dataset revolves around the three types of Penguin species – Adelie, Chinstrap and Gentoo. The study was conducted in the region Anvers on three different islands – Torgersen, Biscoe and Dream. The study aims at taking observations on penguin body features like culmen length / depth (bill length / depth), flipper, BMI etc. The study has a limited number of readings great for some quick data science studies and visualizations.

**DATA CLEANING**

I have done a number of things to clean the data here:

* Get the statistics about the data - Starting with getting more information about the dataset using the describe and info methods. This clearly showed that there are some columns which have got null values. It also gives you an overview of the data like min, max, mean and the datatypes
* Null values - Find out whether there are any null values. For this dataset there are a handful of null values found. This is a very tiny percentage of the overall data and hence I have decided to drop the rows with null values
* Handle the substitute values – There is a single record in the table for which Sex column had value of ‘.’. I had to drop this value to achieve the consistency
* Handling duplicates – I checked for the duplicates in the dataset but there was none found
* Readability: The column Species had too many words in them to describe a type of penguin. This column is now transformed so that a single word recognizes the Penguin species

**MISSING DATA**

There were a handful of records found to be missing across various columns. As this percentage is very tiny compared to the overall dataset, I decided to drop such rows.

**DATA STORIES AND VISUALISATIONS**

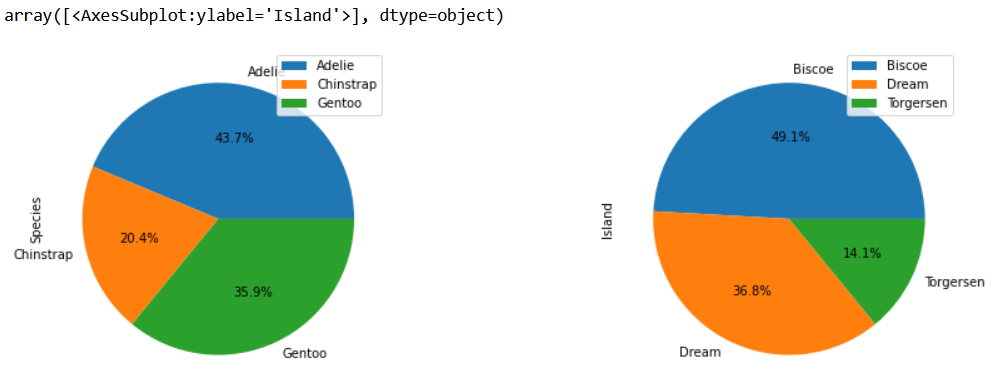
**# Analysis 1**

**## Volume Analysis**

The piecharts below show:

1. Adelie Penguin has the highest records in the dataset with 43.7% where as Chinstrap penguin have the lowest records

2. The highest penguin dataset is taken from Biscoe island with 49% where as Torgersen has the lowest dataset 14.1%



**# Analysis 2**

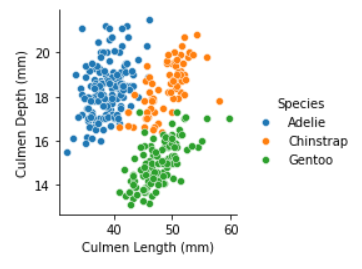
**## Facetgrid - Culment Depth vs Culmen Length By Species**

Below facetgrid shows the culmen depth and length analysis for each type of the Species.

1. It shows that Adelie penguins have got the maximum culmen but they seem to have the least culmen length

2. Gentoo penguins seem to have a the least culmen depth but have good culmen length

3. Chinstrap have got a good average culmen length as well as depth



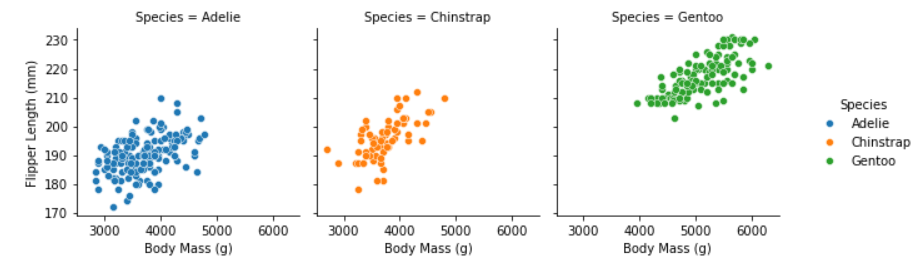
**# Analysis 3**

**## Facetgrid - Body Mass vs Flipper Length by Species**

Please refer to the below facetgrid analysis:

1. It shows Gentoo penguins have got the highest Body mass as well as flipper lengths

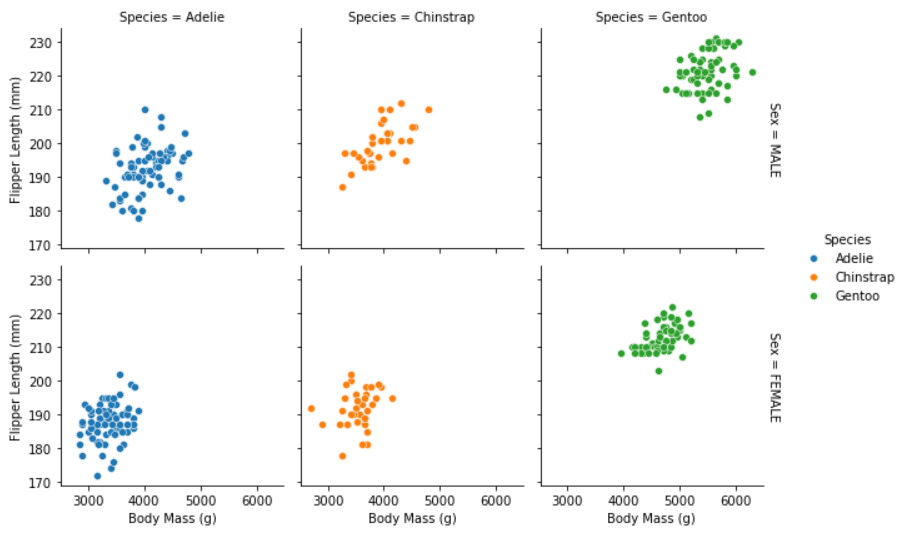
2. Chinstrap and Adelie penguins have got almost the same body mass and flipper lengths



**# Analysis 4**

**## Facetgrid - Flipper length and BMI analysis by Species and Sex**

Male penguins generally have got longer flippers and higher BMIs. This is true across all three species



**# Analysis 5**

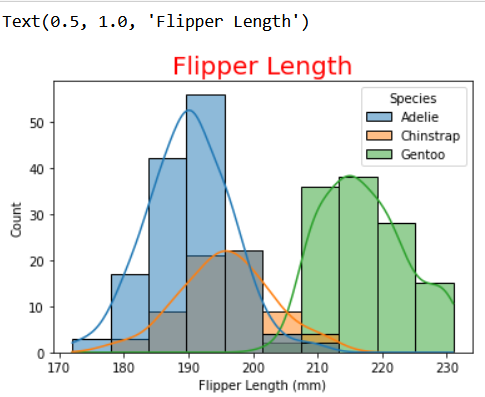
**## Histplot analysis for Flipper Length by Species**

The histplot below shows:

1. Majority of the Adelie penguins have flipper lengths between 180-200mm

2. Majority of the Chinstrap penguins have flippe lengths between 190 and 200mm

3. The Gentoo penguins have got the longest flippers with majority of them peaking beween 205 and 225mm



**# Analysis 6**

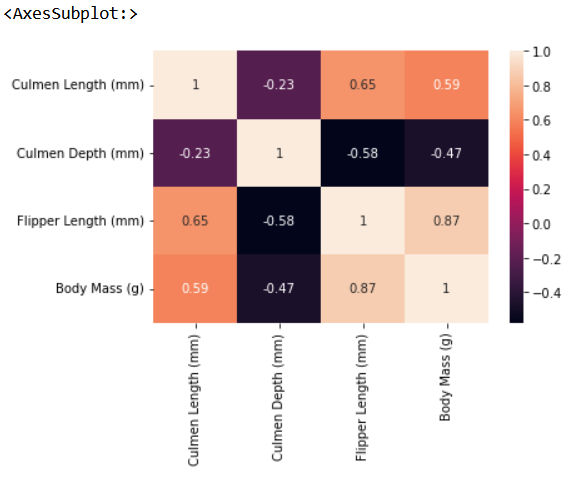
**## Heatmap**

Please refer to the heatmap depicting the correlation between the numeric features

1. Culmen length and depth have got weak to moderate correlation meaning when one goes up the other one reduces

2. Flipper length and Culmen Length have got strong positive correlation, meaning when flipper lengths go up the culmen lengths also go up.

3. Body Mass has got moderate to strong positive correlation with Culmen length but a moderate correlation with Culmen width.



**# Analysis 7**

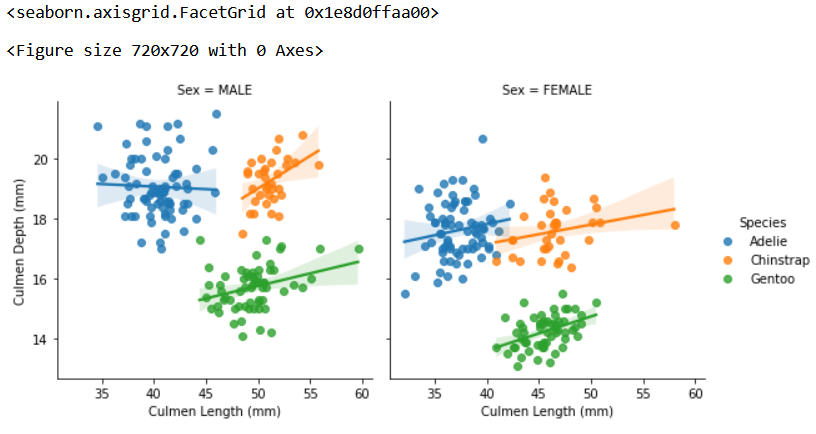
**## lmplot between Culmen Length and Depth per Species**

Looking at the below lmplot, we can see the following:

1. Male have generally got longer and deeper culmen compared to their counter parts female pengune across all the penguins

2. Chinstrap and Gentoo penguins have got weak to moderate correlation between culmen length and depth in both the genders.

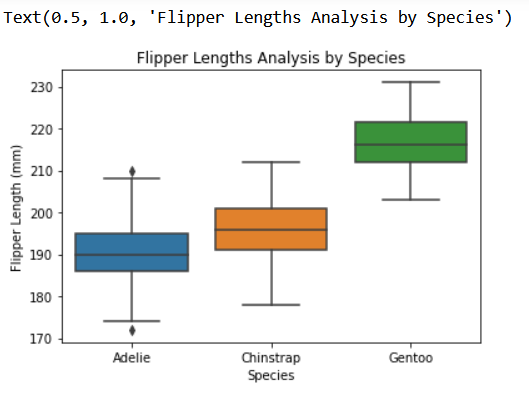
3. Adelie Female penguins have weak to moderate correlation between culment length and depth where as their male counter parts we cannot say the same



**# Analysis 8**

**## Boxplot on Flipper Lengths per Species**

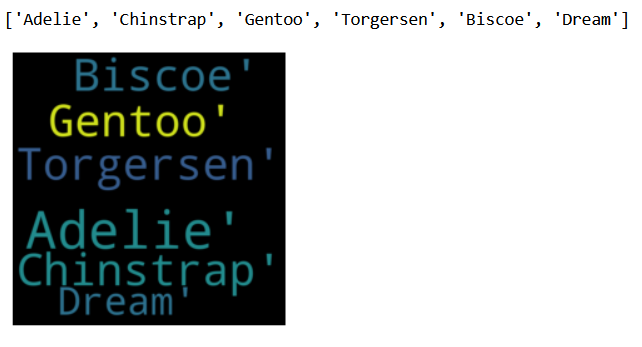
The below barplot gives us fair idea of the flipper lengths across the three penguin species. e.g. The Flipper lengths for Adelie are: Min - 175mm, Max - 208mm and Mean - 190mm. Similarly, we have derive the numbers for other species as well.



**# Analysis 9**

**## WordCloud for Species and Island names**

Using the external library for plotting WordCloud. In this example I have used the names of the species and also Island



**THIS REPORT WAS WRITTEN BY : Maitri Mahida**

