**Statistical Foundations of Data Science (MAS8403)**

**Project**

**Dataset**: PalmerPenguins.csv **Student Name**: Akash Mohandoss

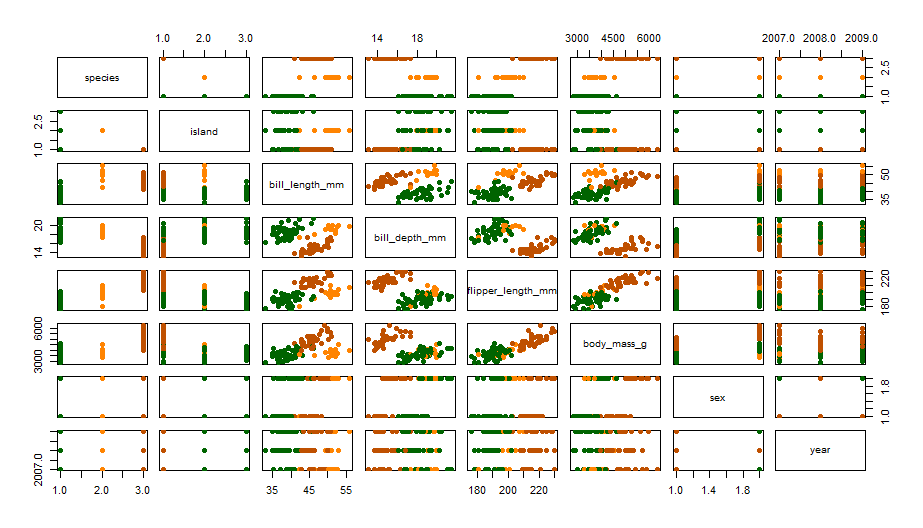
**Student ID**: 220488118

1.Introduction:

The objective of this project is to create a report using the palmerpenguins.csv dataset from palmer archipelago, which estimates the probabilities/ proportion of penguin population from the given dataset in general, determining the sex of the penguin species from the given measurement dataset which researchers at palmer archipelago are finding it hard to predict their sex using traditional method with physical contact and showing how physical characteristics of the penguins differ between islands from the recorded.

The dataset contains data of 100 penguins with 8 variables (species, island, bill\_length\_mm, bill\_depth\_mm, flipper\_length\_mm, body\_mass\_g, sex, year) describing their characteristics and with reference to these variables the insights are identified in this project which clearly answers the questions put forward for this project.

2. Dataset Analysis:

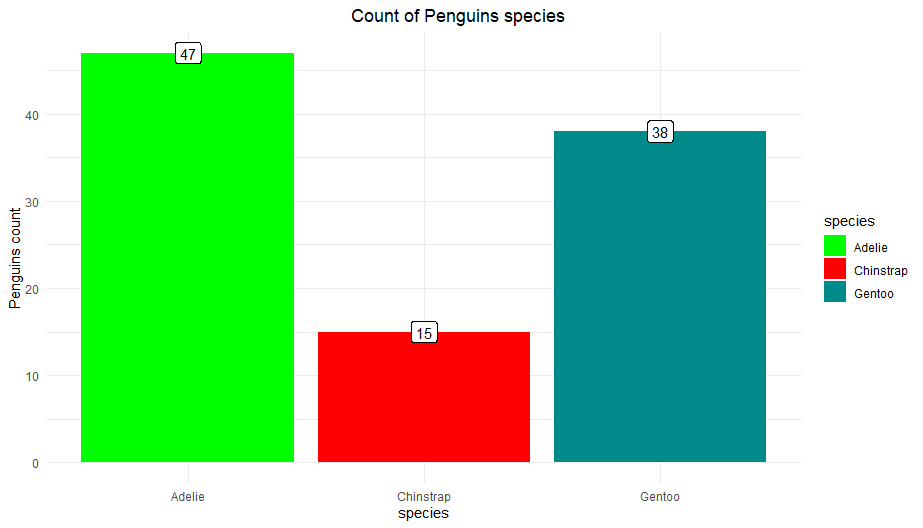
From the given palmerpenguin dataset, Exploratory Data Analysis has been applied to gain useful insights which tells us various characteristics of penguin species and conveying useful information from the dataset.

[Plot 2.1]

The plot 2.1 is a summarization of dataset of species type by all other categorical variables and numerical variable. From the given dataset the variable species, sex, island are the categorical variable and rest of the variables belongs to numerical. The above plot describes how the species differ with their Bill length, bill depth, Flipper length and so on. I have specified colour each species type which makes it clearly understandable.

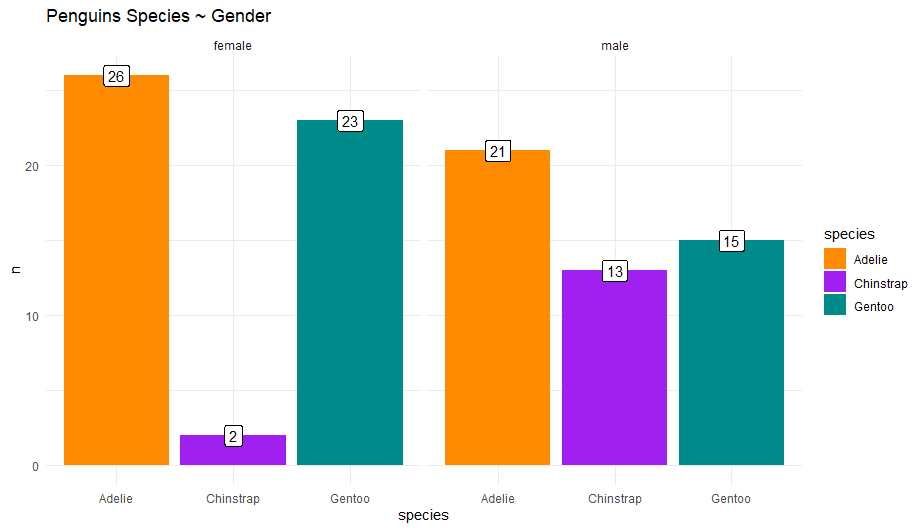
**2.1 Estimating Species Count**

The below plot 2.2 describes the count of penguin’s species namely Adelie, Chinstrap and Gentoo from the data given. It is clearly shown that the population of Adelie is 47, Chinstrap is 15 and Gentoo is 38. This shows us that the out of 100 penguins data Adelie species is the highest population and chinstrap with the least population provided. Further analysis on this penguin’s data will be based on that.



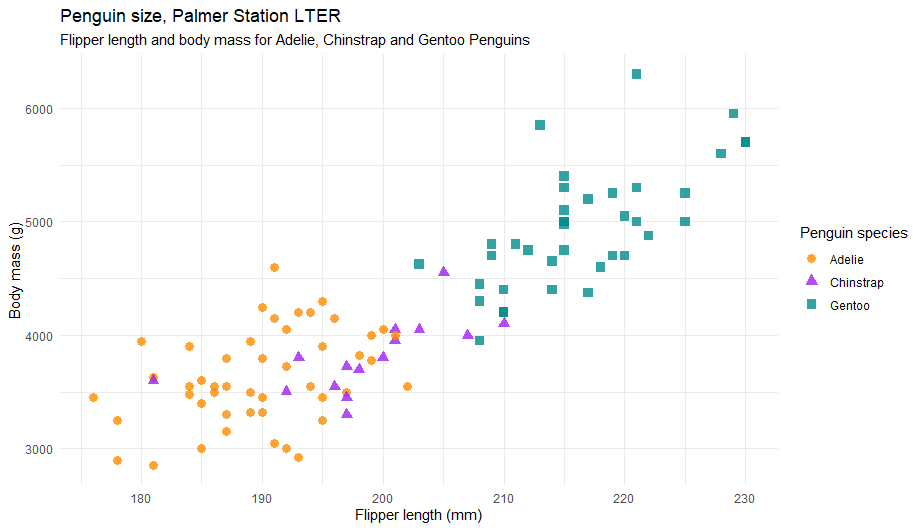
[Plot 2.2]

**2.2 Predicting Penguin Species Count by Sex**



[Plot 2.3]

Plot 2.3 represents the count of respective species and is separated based on their sex. The x-label indicates the type of species and the y-label ‘n’ represents the count of the species. From the bar graph it can be seen that the Adelie population in both the sex are higher than the other species population having 26 in Female and 21 in male. The chinstrap population is the lowest with Male count being 15 and female being 23. So, the population order goes by Adelie > Gentoo > Chinstrap.

**2.3 Predicting Penguins size by body mass and flipper length of penguin**

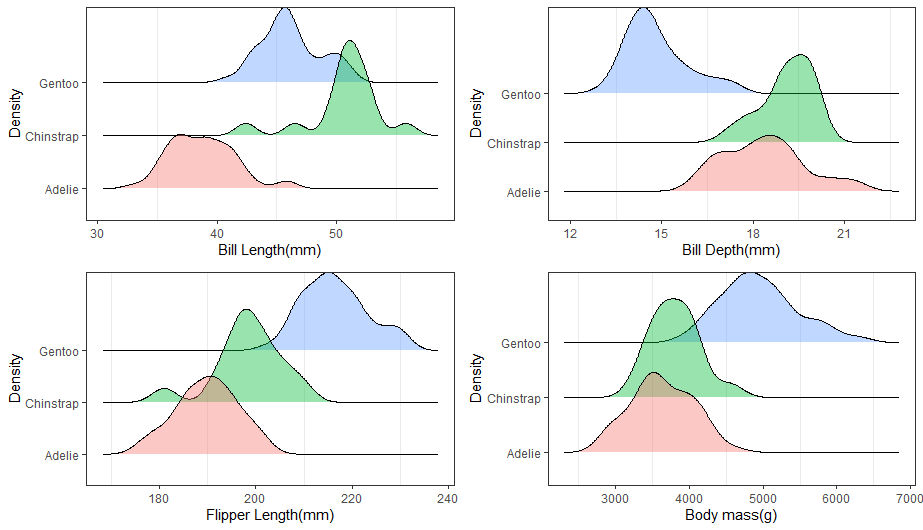
[Plot 2.4]

Comparing the Flipper length with Body mass of the penguin’s, it is clearly shown that the Gentoo species has the highest flipper length and body mass. From Plot 2.4, we can infer that as the flipper length increases and the body mass is increased the Gentoo species are alone recorded. Whereas if we consider Adelie species, they are the smallest compared to the other species. The chinstrap species have an average flipper length among other species which means they are neither small nor large.

The body masses of Adelie and Chinstrap species are almost same which can easily be inferred from the plot above. So, from the above consideration I conclude that Gentoo species are the ones to be the giants among other species. They have the highest flipper length and body masses when compared to other species in Palmer Archipelago.

**2.4 Density Plots of Numerical variable with Species (**Flipper length, bill length, bill depth, body mass**)**

Density distribution plot has been plotted with species against numerical variable. The first plot against bill length



Density Plots of Numerical variable

[Plot 2.5]