Palmer Penguins Analysis

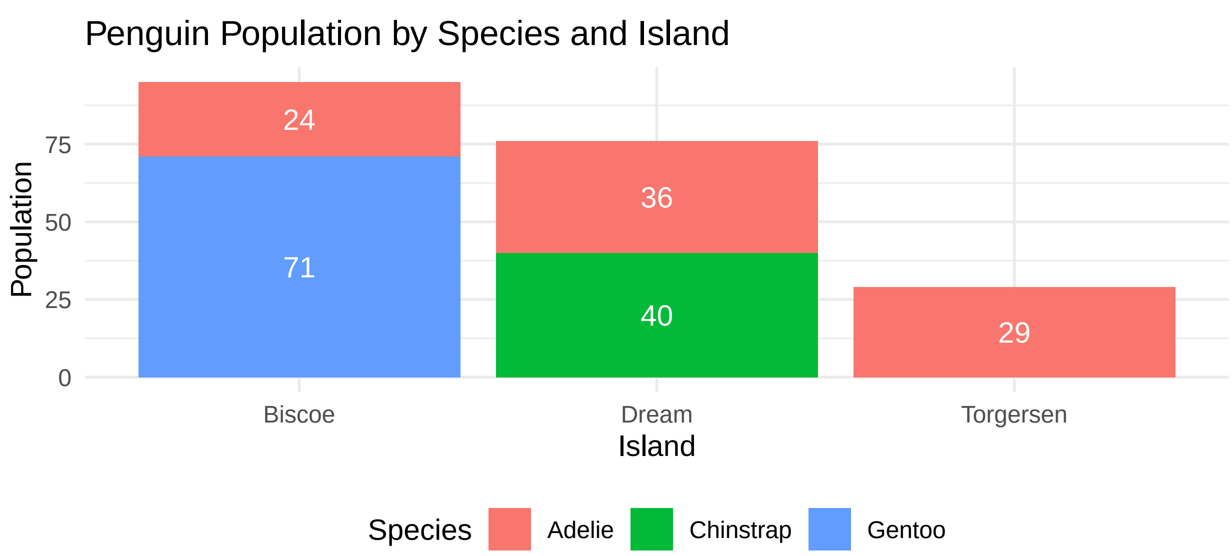
Harish Sainath Sai Prasath (230164781)

Introduction

In this report, we aim to analyse a subset of the famous *Palmer Penguins* data set. The subset data used for analysis has 200 entries and 8 different variables describing the penguins. These variables are *species, island, bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g, sex, year* . Before diving deep into the data, let's try to get a better understanding about the penguins given in the data set by exploring all the variables, one at a time.

Species and Islands:

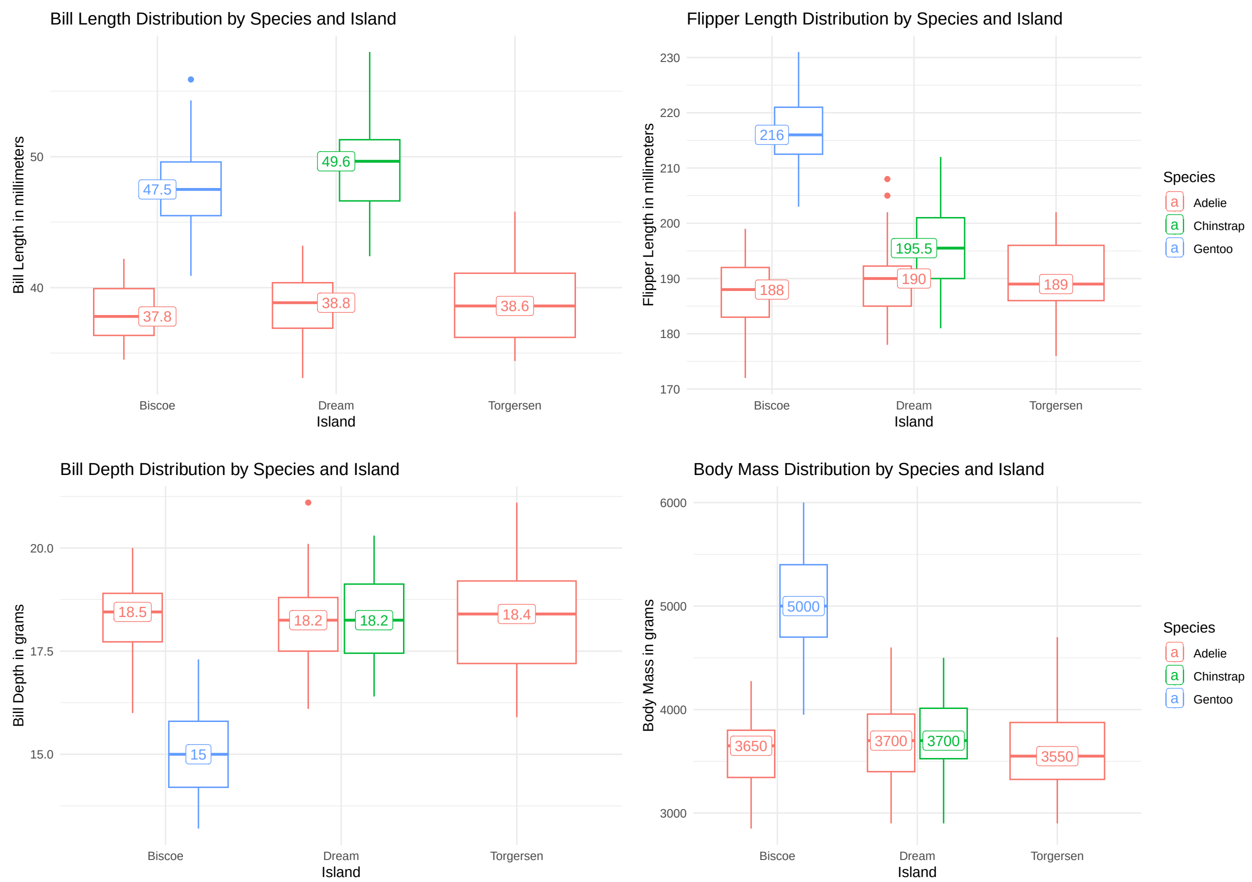
- There are 3 different islands and these are *Biscoe, Dream, Torgersen.*
- We can observe from the below plot that, Biscoe is the most populated island followed by Dream and Torgersen.
- There are 3 species of penguins and they are *Adelie, Chinstrap, Gentoo* respectively.
- Population of each species in these island is given in the below plot.



Bill Length, Bill Depth, Flipper Length and Body Mass:

- To get an idea about the features of penguins, we will consider median as a the measure of central tendency because it is not affected if the data is skewed.
- The penguins have a median bill length of *43.5 mm.*
- From the below box plots we can observe that Chinstrap Penguins has the highest median Bill length. This is closely followed by Gentoo penguins.

- Apart from the median bill lengths, a few interesting points can also be observed. Adelie is the most common type of penguin that is present in all three islands. Chinstrap is present only in Dream island, and Gentoo is present only present in Biscoe island.
- All penguins have a median bill depth of *17.3 mm*. Median bill depth for each species can be observed from the below box plots.
- Gentoo Penguins have the highest median flipper length and Adelie has the overall lowest median flipper length.
- The penguins have a median body mass of *3950 g*.
- Gentoo penguins are the heaviest penguins in the dataset. Adelie penguins are the the least heavy of the group and chinstrap penguins are comparable to Adelie as per body mass is concerned.

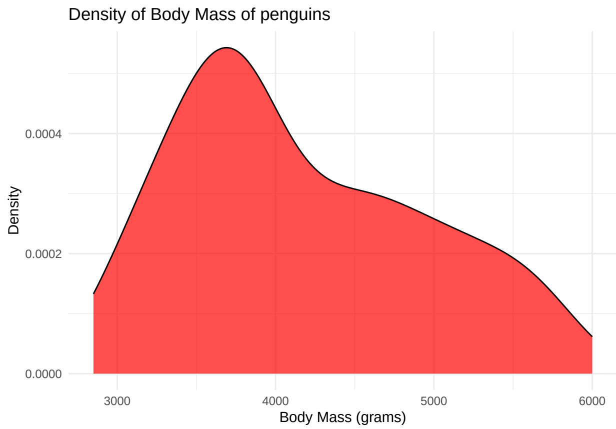


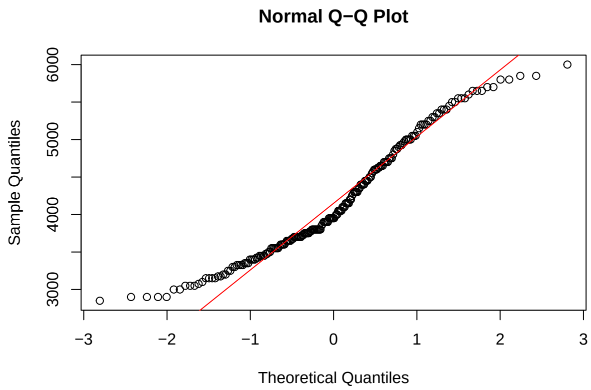
Year and Sex :

- The data is collected for 3 years and these are *2007, 2008, 2009* respectively. There are 98 female penguins and 102 male penguins in the dataset.

Distribution Fitting and Population Proportion Estimation

Let's take one Body mass of penguins and fit a distribution for proportion estimation. For getting initial impressions of how the data is distributed, let's plot a density plot. Also a lot of things in nature are observed to be falling under a normal distribution. So Let's use a q-q plot to check if body mass can be fitted using a Normal/Gaussian distribution.





Two major things can be observed the above plots.

- q-q plot shows that fitting a normal distribution might not be the best choice for good parameter estimations.
- Distribution of body mass seems to right skewed.

Now we have to look for some other distribution to fit the data. The body mass is a positive value and the data is right skewed, *LogNormal Distribution* might be a good choice to fit the data.

so let's fit LogNormal Distribution to our data.

Fitting the Data to a LogNormal Distribution

Probability density function for a LogNormal distribution is given by

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{\left[-\frac{(ln(x)-\mu)^2}{2\sigma^2}\right]}$$

- μ is the mean of the normal distribution. - σ is the standard deviation of the normal distribution. - n is equal to 200 in our case.

Likelihood function:

$$L(\mu, \sigma|x) = \sum_{i=1}^{n} \ln(f(x_i|\mu, \sigma))$$

Log Likelihood function:

$$L(\mu, \sigma|x_n) = \sum_{i=1}^{n} \left(-\ln(x_i\sigma\sqrt{2\pi}) - \frac{(\ln(x_i) - \mu)^2}{2\sigma^2} \right)$$

$$L(\mu, \sigma|x_n) = -n\ln(\sigma) - n\ln(\sqrt{2\pi}) - \sum_{i=1}^{n} \ln(x_i) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\ln(x_i) - \mu)^2$$

Now using Partial differentiation we find equations for μ and σ.

- To estimate parameter μ (Mean) we differentiate the log-likelihood with respect to μ

$$\frac{\partial L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(\ln(x_i) - \mu) = 0$$

Solving this equation we get,

$$\text{mean: } \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\ln(x_i)$$

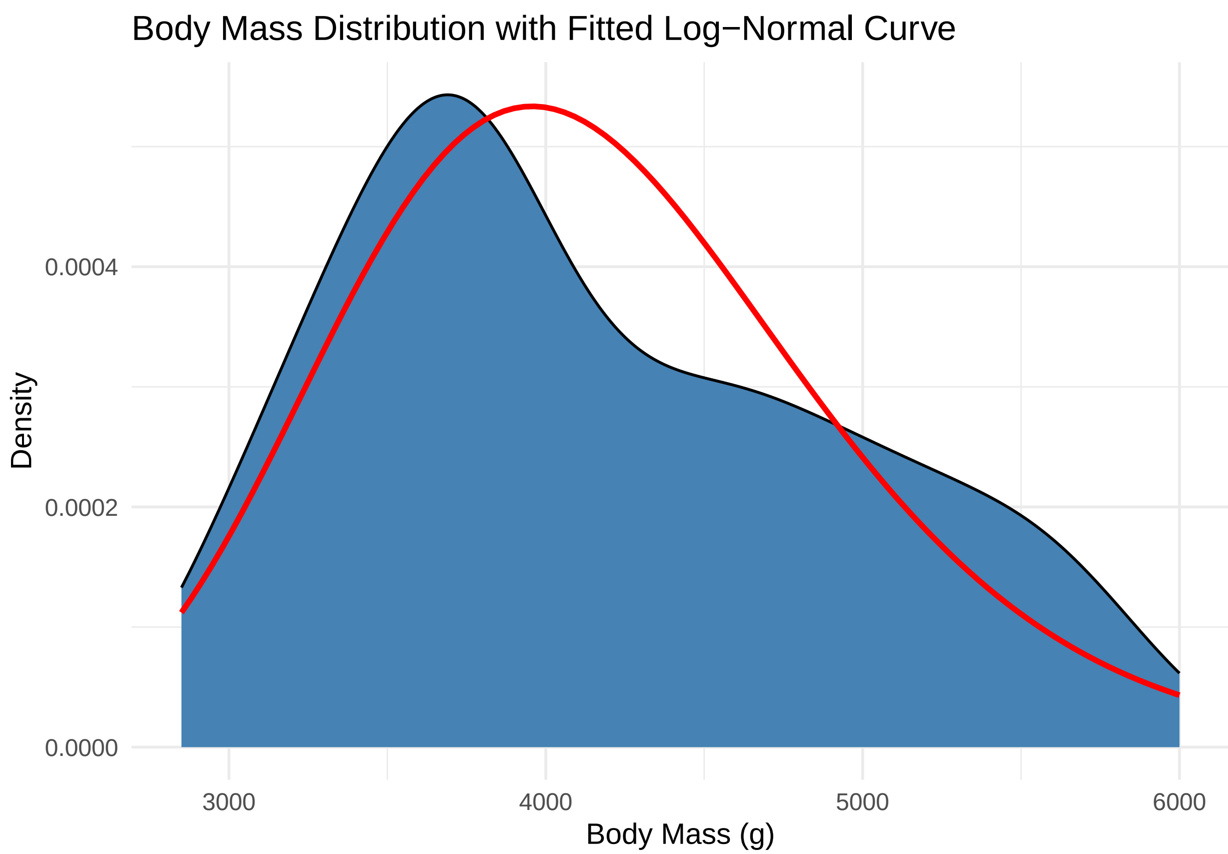
- To estimate parameter for σ^2 (Variance) we differentiate the log-likelihood with respect to σ^2

$$\frac{\partial L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(\ln(x_i) - \mu)^2 = 0$$

by solving this equation, we get

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\ln(x_i) - \hat{\mu})^2$$

$$\text{Standard deviation: } \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\ln(x_i) - \hat{\mu})^2}$$



Limitations of this Model:

- While this LogNormal represents skewness, it still does not match the actual data. But this is far better than Normal distribution where skewness is zero.