MAS8403 Final Assignment (Student No. : 230404311)

Hrithik Kumar

2023-10-17

Overview

In this report, we analyse a dataset containing information about different species of penguins located in The Palmer Station, Antarctica. The data is present in the palmerpenguins library.

The data set contains 200 rows and 8 columns. Each row represents information about the species of Penguin(Adelie, Gentoo, Chinstrap), the island it is from(Torgersen, Biscoe, Dream), its gender and some body measurements. Let's try to understand the distribution of the data for different species of penguins and the islands they belong to.

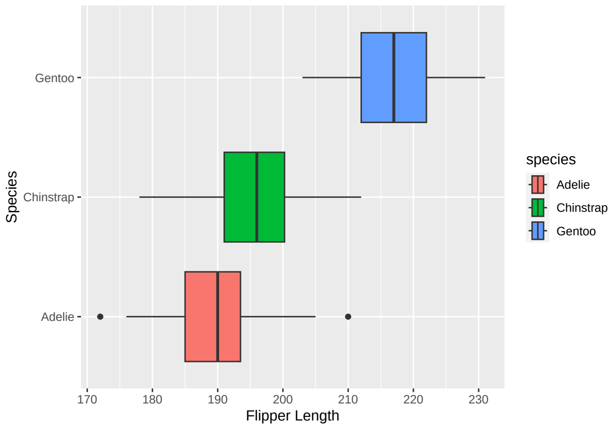
```
##
##              Biscoe Dream Torgersen
##   Adelie         28    26        29
##   Chinstrap       0    52         0
##   Gentoo         65     0         0
```

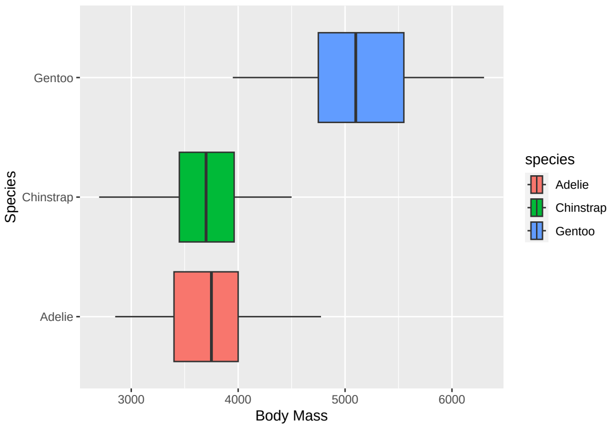
```
##     species flipper_length_mm bill_length_mm bill_depth_mm body_mass_g
## 1    Adelie          189.6506       39.11084      18.34819    3739.458
## 2 Chinstrap          195.3654       49.18077      18.45962    3719.712
## 3    Gentoo          217.3846       47.82308      15.02000    5133.077
```

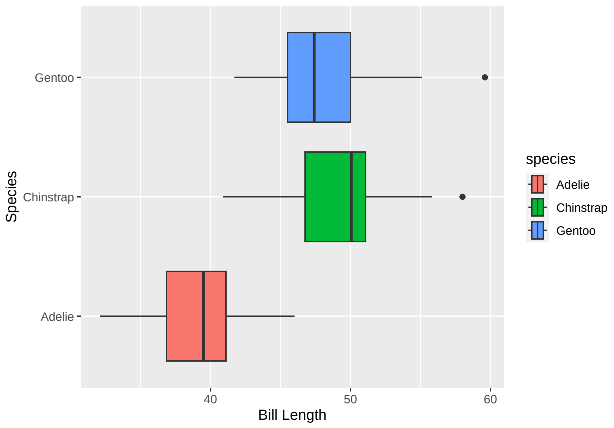
We have roughly 50-60 samples from each species. If we consider the average values of the body measurements among the species, we can see clear trends such as Gentoo penguins have longer flippers and weigh more than Adelie and Chinstrap penguins. They have slightly less bill depth compared to others. However, these averages don't represent the distribution of the data as a whole and thus a boxplot will be useful in understanding the distribution of data and see outliers, if there are any.

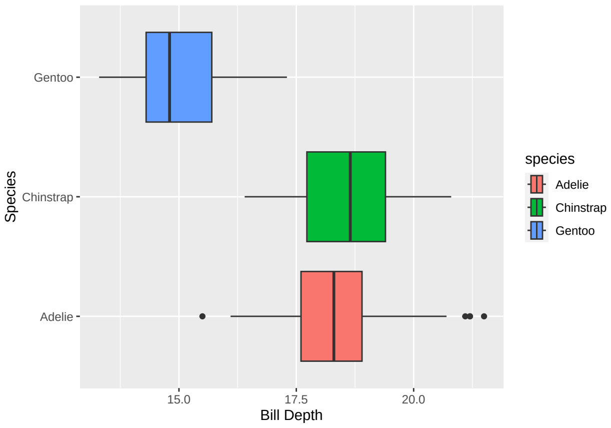
Body Measurements

We have enough information from each species of penguins to further proceed with our analysis. We will now try to compare different penguins based on their body measurements such as bill length, bill depth, flipper length and body mass. Note that the lengths are calculated in mm and the body mass in grams.





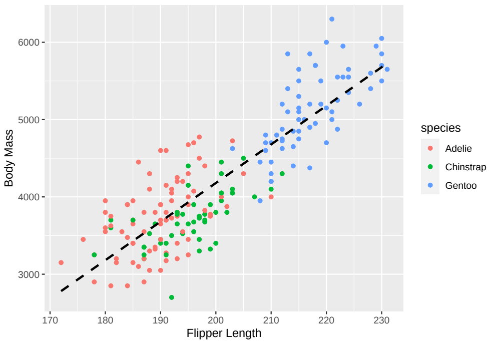


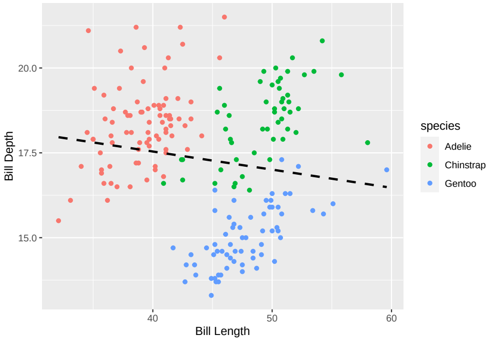


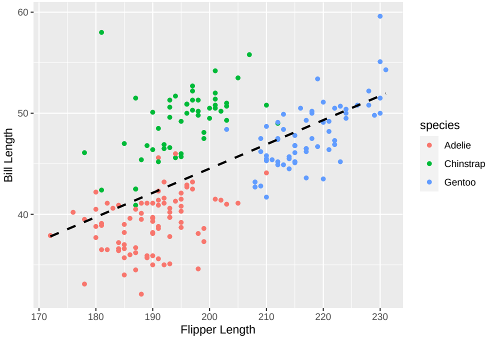
The comparison of Flipper length suggests that Adelie and Chinstrap penguins have flipper length in the same range, however, for Gentoo penguins, the range is slightly higher. Similarly, Bill depth and Body mass of Adelie and Chinstrap penguins are roughly in the same range, and for Gentoo penguins the range is lower for bill depth and higher for body mass. This confirms our initial observation regarding flipper length and bill depth. Finally, the Bill length of Chinstrap and Gentoo penguins is similar in range with median bill length of Chinstrap being higher, but Adelie penguins have slightly smaller Bills in length.

Relationship between Features

Now that we can differentiate between different species of penguins based on their features, let's try to look a little bit deeper and find any relations between different features. Since, these are all quantitative variables, Scatter plot would be best choice for comparison and discovering underlying trends in the data set.

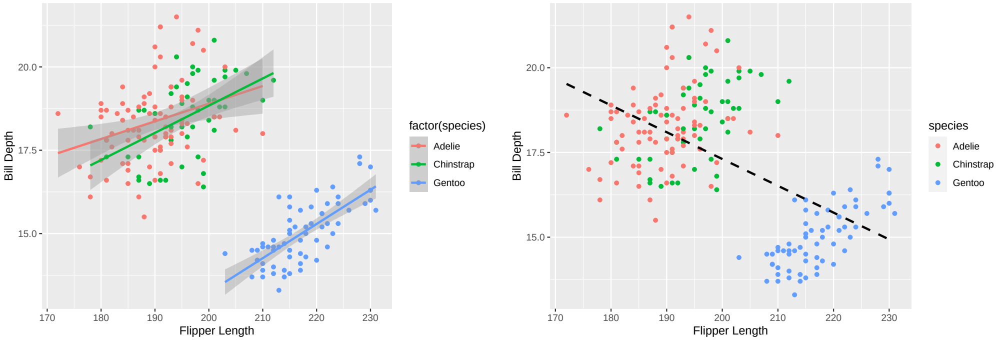






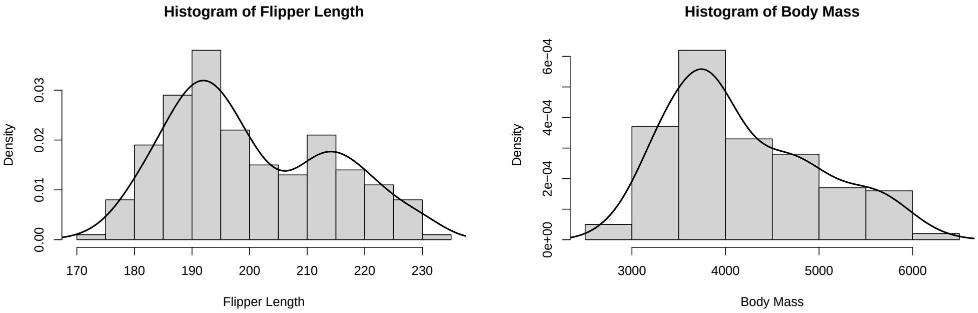
The scatter plot of flipper length vs Body mass shows a direct relation between the two. For all 3 species, as flipper length increases, body mass increases as well. In the scatter plot of flipper length vs. bill length, we can observe similar trends, albeit the correlation is not as readily discernible as the plot of flipper length vs Body mass.

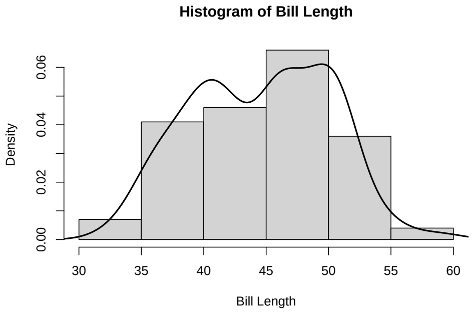
It is important to note that we see these scatter plots as a whole and not look at each species individually. If we consider the scatter plot for Flipper Length vs Bill Depth, there is an increasing trend for all 3 species individually. However, the trend does not hold when we look at them together. In fact, there is an observable downward trend suggesting an inverse relationship. This is known as Simpson's Paradox and should be avoided while making any statements regarding a data set containing different sub groups or categories.



Probability Distribution of Continuous Variables

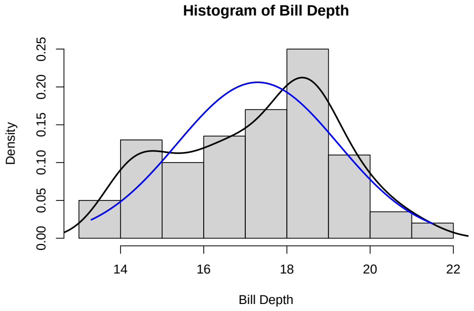
Now that we have done some exploratory analysis on the data, let's try to choose an appropriate probability distribution to represent one of the quantitative variables. We will be using the density plot for each of the variables(flipper_length_mm, body_mass_g, bill_length_mm, bill_depth_mm).



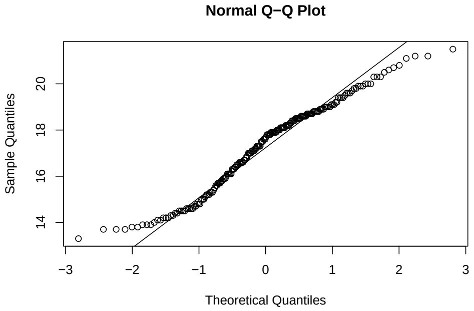


Histogram of Bill Length

Upon closely observing the histogram and density plots, we can say that the distribution of Bill Depth resembles a Normal Distribution. We can further support our decision by plotting Q-Q plot.



Histogram of Bill Depth



Normal Q−Q Plot

As we can see that data points are quiet closely concentrated near the diagonal, which is an indicator of the random variable resembling normal distribution.

Now that we have chosen a probability distribution for Bill Depth, let's derive the Maximum Likelihood function for the parameters to be estimated.

Assuming that the 200 observations made for bill depth can be modeled as realizations of independent and identically distributed(IID) random variables $X_1, \ldots, X_{200}$ where each $X_i \sim \mathrm{N}(\mu, \sigma^2)$, let's consider the Probability density function and take the logarithm.

$$
\begin{aligned}
p(x|\mu) &= \frac{e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}}{\sigma\sqrt{2\pi}} \\
\log p(x|\mu) &= \log\left(\frac{e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}}{\sigma\sqrt{2\pi}}\right) \\
&= -\frac{1}{2}(\frac{x-\mu}{\sigma})^2 - \log\sigma\sqrt{2\pi}
\end{aligned}
$$

Now, we calculate the log-likelihood function

$$
\begin{aligned}
\ell(\mu|\boldsymbol{x}) &= \log p(x_1|\mu) + \log p(x_2|\mu) + \log p(x_3|\mu) + \log p(x_4|\mu) + \log p(x_5|\mu) + \ldots + \log p(x_{200}|\mu) \\
&= -\frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2 - \log\sigma\sqrt{2\pi} - \frac{1}{2}\left(\frac{x_2-\mu}{\sigma}\right)^2 - \log\sigma\sqrt{2\pi} - \frac{1}{2}\left(\frac{x_3-\mu}{\sigma}\right)^2 - \log\sigma\sqrt{2\pi} \\
&\quad - \frac{1}{2}\left(\frac{x_4-\mu}{\sigma}\right)^2 - \log\sigma\sqrt{2\pi} - \frac{1}{2}\left(\frac{x_5-\mu}{\sigma}\right)^2 - \log\sigma\sqrt{2\pi} - \ldots \\
&= \left(-\sum_{i=1}^{200}(x_i-\mu)^2\right)\frac{1}{2\sigma^2} - K
\end{aligned}
$$

where $K = 200 \log \sigma \sqrt{2\pi}$ is a constant which does not depend on μ.

To get the MLE, we differentiate the log-likelihood function with respect to μ

$$\begin{aligned}
\frac{\partial}{\partial \mu} l(\mu|\boldsymbol{x}) &= \frac{\partial}{\partial \mu}\left[\left(-\sum_{i=1}^{200}(x_i-\mu)^2\right)\ \frac{1}{2\sigma^2} - K\right] \\
&= -\frac{1}{2\sigma^2}\left(\frac{\partial}{\partial \mu}\sum_{i=1}^{200}(x_i-\mu)^2\right) - \frac{\partial}{\partial \mu}K \\
&= -\frac{1}{2\sigma^2}\left(\sum_{i=1}^{200}2(x_i-\mu)(-1)\right) \\
&= \left(\sum_{i=1}^{200}x_i - \hat{\mu}\right) \times \frac{1}{\sigma^2}
\end{aligned}$$

Now we solve for $\hat{\mu}$ by equating the above result to 0.

$$\begin{aligned}
\left(\sum_{i=1}^{200}x_i - \hat{\mu}\right) \times \frac{1}{\sigma^2} &= 0 \\
\left(\sum_{i=1}^{200}x_i - \hat{\mu}\right) &= 0 \\
\sum_{i=1}^{200}x_i &= 200\hat{\mu} \\
\therefore \qquad \hat{\mu} &= \frac{1}{200}\sum_{i=1}^{200}x_i
\end{aligned}$$

Similarly, we can derive the MLE for σ^2 as well.

$$\hat{\sigma}^2 = \frac{1}{200}\sum_{i=1}^{200}(x_i - \hat{\mu})^2$$

Now that we have MLE functions for the parameters, we can calculate Mean and Variance for Bill Depth.

```
## Mean :  17.2955
```

```
##
## Variance :  3.74606
```

These values are quite helpful, however, these are not sufficient to make inferences about the population. There is a certain level of uncertainty with these values and this uncertainty can be quantified with the help of Confidence Intervals.

To calculate Confidence Intervals of a normal distribution, we use Student's t-distribution formula for population mean.

$$\bar{x} - t_{n-1,1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{n-1,1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$$

```r
alpha <- 1 - 0.05
t_val <- qt(1-alpha/2,199) #199 because n = 200 for dataset

lower_bound <-
```

```
  mean(my.penguins$bill_depth_mm) - t_val * sd(my.penguins$bill_depth_mm)/sqrt(200)

upper_bound <-
  mean(my.penguins$bill_depth_mm) + t_val * sd(my.penguins$bill_depth_mm)/sqrt(200)
```

```
## Confidence Interval: ( 17.28691 ,  17.30409 )
```

We calculated the 95% confidence interval which is between 17.2869 and 17.3041. What this really means is that if samples were collected from the population over and over then, 95% of the times the population mean would lie in the above interval. This is much better than having just one value as representation of a whole population or sample and takes into account the uncertainty that comes with making inferences based on a sample.

Based on above calculations for estimates of parameters and comparisons of histogram, density and Q-Q plots, we can say that Normal distribution is a good fit for the data. However, its not the only distribution suitable for the data. Most real world data can fit multiple distributions. The histogram of Bill depth suggests slight asymmetry in the data which can be a reason to not use normal distribution. Moreover, normal distribution doesn't work well when there are outliers present in the data and we do have outliers present as pointed out earlier by the box plot of Bill depth. Even though normal distribution is quite helpful in estimations and modelling real world data, it does not hold well when data is skewed or is tail heavy, hence, other distributions such poisson or log normal distributions are used as an alternative.

Variables suitable for Determining Sex

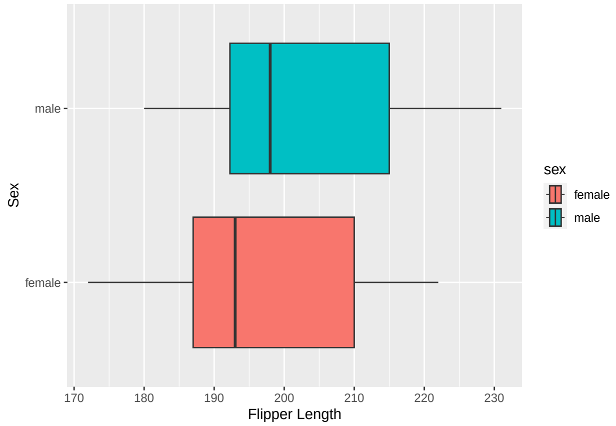
So far we have discussed about the data set from the perspective of being able to differentiate between different species of Penguins. Let's try to explore the data so that we can differentiate the gender of the Penguin.

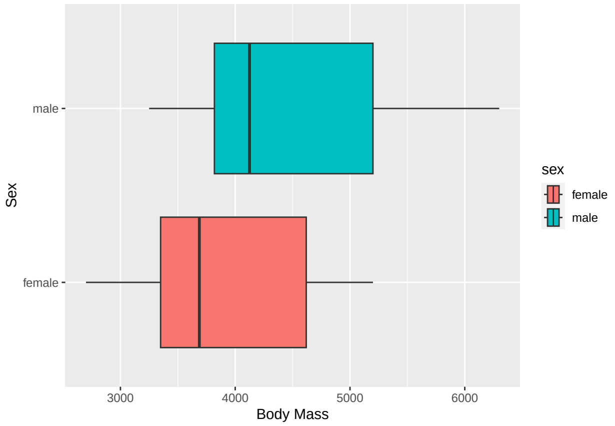
The average values of all the body measurements is a good starting point to begin our exploration.

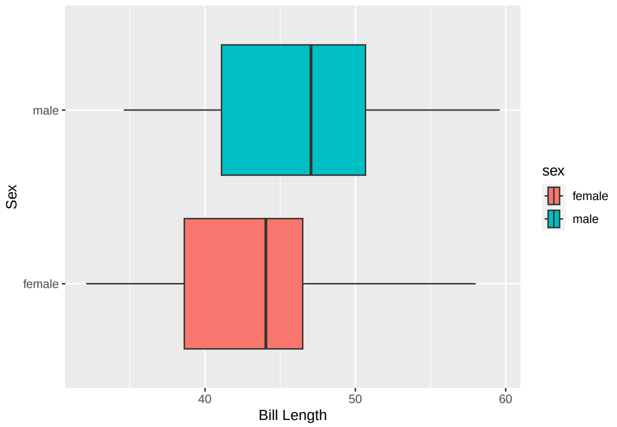
```
##      sex flipper_length_mm bill_length_mm bill_depth_mm body_mass_g
## 1 female          197.1277       42.84149      16.41064    3874.468
## 2   male          202.8302       46.08491      18.08019    4464.623
```

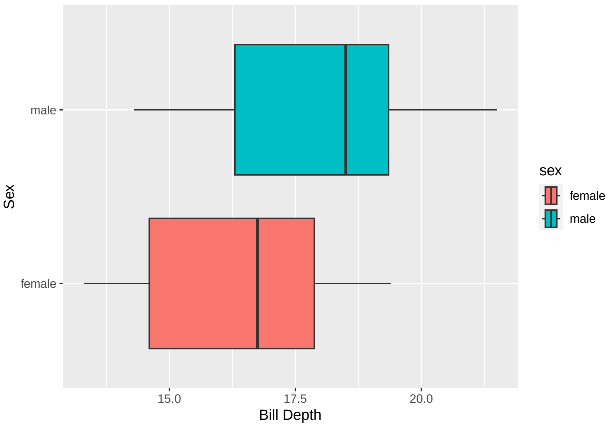
Initial observations tell us that male penguins have longer flippers on average and thus also have high body mass. We have already observed this while looking for trends among different variable relationships for different species. Bill Length and depth of male penguins is also higher.

These values will make more sense if we plot a box plot for these random variables. Box plot will also display the minimum and maximum of these variables and display outliers, if any.









Most observable difference between male and female penguins is shown by bill depth. Range of bill depth for male penguins is quite higher than female penguins. This can form the basis of our hypothesis and we can assess this using Two sample t-test to determine how consistent it is with the data.

Since we have already considered that Bill Depth follows normal distribution, we can go ahead and form our hypothesis.

Let Null Hypothesis H_0 be that there is no significant difference in the bill depth between male and female penguins and Alternative Hypothesis H_1 be that there is a significant difference in the bill depth of male and female penguins.

$$H_0 : \mu_{male} = \mu_{female} \quad and \quad H_1 : \mu_{male} \neq \mu_{female}$$

In order to test the above hypothesis, we first need to deal with Variances. We will be performing Bartlett's test to determine whether the variances of the two populations are equal or not.

```
##
##  Bartlett test of homogeneity of variances
##
## data:  list(bill_depth_male, bill_depth_female)
## Bartlett's K-squared = 0.058791, df = 1, p-value = 0.8084
```

We get the p-value of 0.8084, which is greater than 0.05, so the assumption of equal variances is valid for our t-test.

Now, with this assumption we move on to compare the means of Bill Depth with respect to gender of the penguins.

```
##
##  Two Sample t-test
##
## data:  bill_depth_mm by sex
## t = -6.7326, df = 198, p-value = 1.761e-10
## alternative hypothesis: true difference in means between group female and group male is not equal to
## 95 percent confidence interval:
##  -2.15857 -1.18053
## sample estimates:
## mean in group female   mean in group male
##             16.41064             18.08019
```

We get p-value of 0.0000000001761, which is less 0.001. This suggests that there is strong evidence against H_0. As data is not consistent with H_0, we'll reject it and go with H_1.

Our Alternative Hypothesis stated that there is a significant difference between bill depth of male and female penguins. Using mean bill depth of male and female penguins, the p-value of Two Sample t-test supports the same idea. Therefore, Bill depth would be a great deciding factor for differentiating male and female penguins.

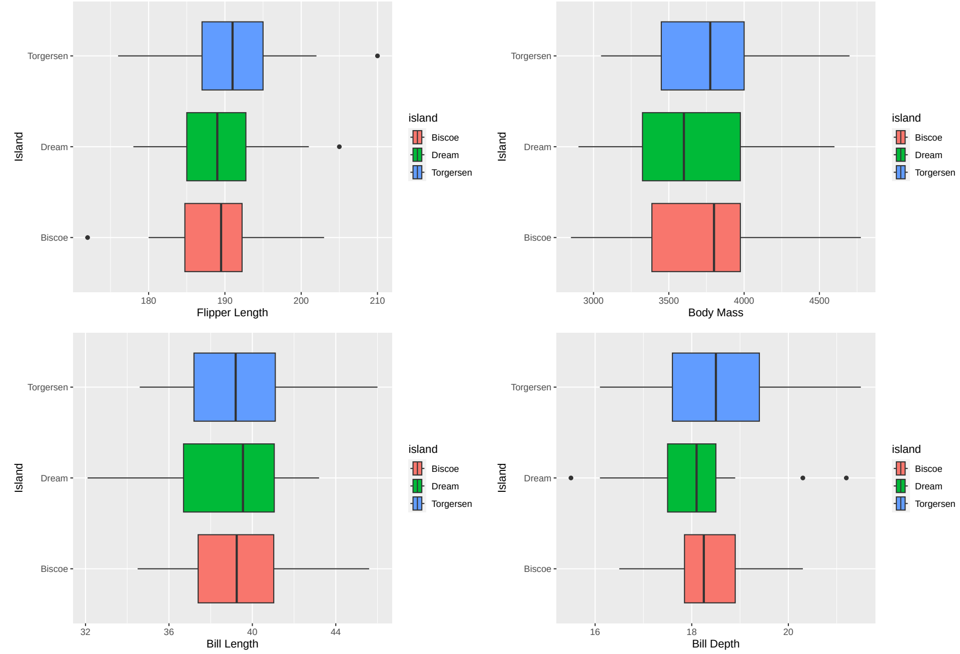
Physical Characteristics based on Islands

Finally, let's try to explore the idea of body measurements correlation with the island the penguins are from.

```
##
##             Biscoe Dream Torgersen
##   Adelie        28    26        29
##   Chinstrap      0    52         0
##   Gentoo        65     0         0
```

The above table provides quite useful information. Firstly, only Adelie penguins are present on all 3 islands. Chinstrap and Gentoo penguins are not present on all 3 islands and thus it is difficult to make any inference about their characteristics being affected by the island they are from.

So, we will be comparing Adelie penguin's characteristics across all 3 islands to determine if their are any trends.



The flipper length and body mass of Adelie penguins is roughly in the same range across all 3 islands. Bill Depth is significantly higher on Torgersen than the other two islands. Bill length and Body mass are also nearly the same across all 3 islands.

Even though these trends might seem interesting, we should keep in mind that these values are representation of a sample of 25-30 observations only. Also, these trends might not be followed by other species of penguins on the same island.

We could go for a t-test to determine whether these differences in values can lead to a usable result but since we have 3 islands i.e 3 Categories, it's not suitable. There are couple of reasons for that, first, t-tests are designed to compare two groups. Second, these tests could lead to Type-I and Type-II errors. And lastly, we could lose context of how all groups compare to each other.