Data Analysis of Palmer Penguins using Statistical Methods

2024-10-18

Pranav Sunil Raja
Student No.: 240408545 Newcastle University, Newcastle upon Tyne

Introduction

This Project aims to apply diverse statistical methods to analyse the Palmer Archipelago Penguin's data-set. This is done by examining the relationship between various factors and islands with the use of various data-set features. The concepts equipped in this project include exploratory data analysis and hypothesis testing, providing essential information and tools for scientists to draw meaningful inferences and make predictive assessments.

Summary of my.penguins

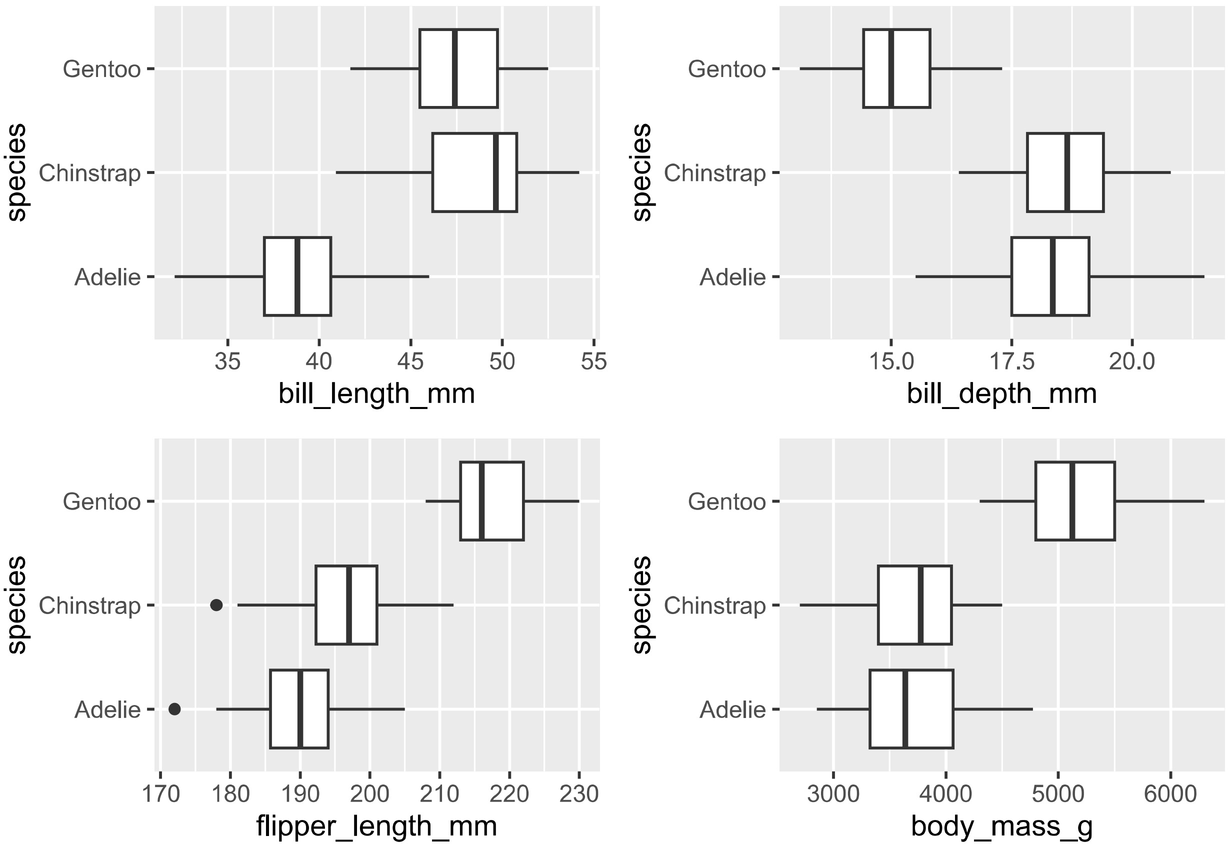
```
summary(my.penguins)
```

```
##       species          island    bill_length_mm  bill_depth_mm
##  Adelie   :92   Biscoe   :95   Min.   :32.10   Min.   :13.10
##  Chinstrap:42   Dream    :75   1st Qu.:39.15   1st Qu.:15.80
##  Gentoo   :66   Torgersen:30   Median :43.35   Median :17.60
##                                Mean   :43.73   Mean   :17.28
##                                3rd Qu.:48.85   3rd Qu.:18.73
##                                Max.   :54.20   Max.   :21.50
##  flipper_length_mm  body_mass_g       sex           year
##  Min.   :172.0     Min.   :2700   female: 97   Min.   :2007
##  1st Qu.:190.0     1st Qu.:3500   male  :103   1st Qu.:2007
##  Median :196.5     Median :4050                Median :2008
##  Mean   :200.3     Mean   :4181                Mean   :2008
##  3rd Qu.:212.2     3rd Qu.:4781                3rd Qu.:2009
##  Max.   :230.0     Max.   :6300                Max.   :2009
```

The dataset of 200 penguins offers a rich information about Antarctica penguin diversity and characteristics. The sample includes 3 samples 3 types of penguins i.e Adelie (92), Chinstrap (42), and Gentoo (66) distributed across 3 islands Biscoe (95), Dream (75), and Torgersen (30). The physical measurements reveal significant variations bill lengths range from 32.1 to 54.2 mm with a median of 43.35 mm, indicating substantial differences maybe related to penguins. Body mass varies widely from 2700 to 6300 g with a median of 4050 g, which could reflect differences in penguins size, age, or health status. Similarly, flipper length ranges from 172 to 230 mm. This flipper length might correlate with swimming efficiency or species-specific adaptations of the penguin. Data set has data spanning from 2007 to 2009 allows for potential analysis of trends and variations. The sample is slightly male-biased (103 males vs. 97 females), which could be important for breeding studies.

1. Exploratory Data Analysis

```
gg1 = ggplot(my.penguins, mapping = aes(bill_length_mm, species)) + geom_boxplot()
gg2 = ggplot(my.penguins, mapping = aes(bill_depth_mm, species)) + geom_boxplot()
gg3 = ggplot(my.penguins, mapping = aes(flipper_length_mm, species)) + geom_boxplot()
gg4 = ggplot(my.penguins, mapping = aes(body_mass_g, species)) + geom_boxplot()
grid.arrange(gg1, gg2, gg3, gg4)
```



The four block plots gives a comparison of key features between 3 penguin species Adelie, Chinstrap, and Gentoo. The graphs signify the distribution of bill length, bill depth, flipper length, and body mass for each species.

1. Gentoo penguins appear to have the longest bills, followed by Chinstrap, with Adelie having the shortest. There's some overlap between Chinstrap and Gentoo, but Adelie is distinctly separate.
2. Adelie penguins show the greatest bill depth, while Gentoo have the shallowest. Chinstrap bills fall between the other two species in depth.
3. Gentoo penguins clearly have the longest flippers, with little overlap with the other species. Chinstrap and Adelie have similar flipper lengths, though Chinstrap flippers appear slightly longer on average.
4. Gentoo penguins are significantly heavier than the other two species. Adelie and Chinstrap have similar body masses, with Chinstrap showing a slightly higher median.

The physical differences between the species are evident in these plots, which emphasize the unique traits of these penguins. These variations may be a sign of species adaptability to various environments or eating habits.

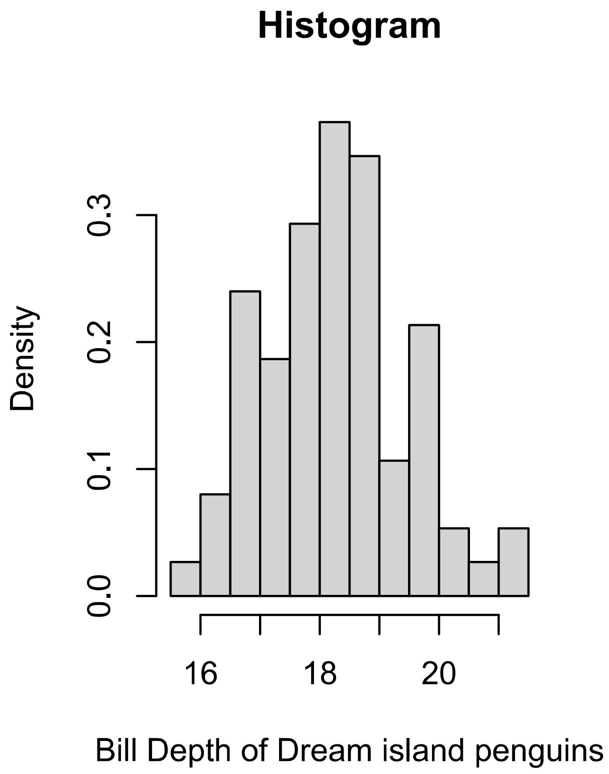
```
table = xtabs(~ year + species + island, data =my.penguins)
table
```

```
## , , island = Biscoe
## 
##       species
## year   Adelie Chinstrap Gentoo
##   2007      6         0     15
##   2008     11         0     23
##   2009     12         0     28
## 
## , , island = Dream
## 
##       species
## year   Adelie Chinstrap Gentoo
##   2007     10        17      0
##   2008      8         9      0
##   2009     15        16      0
## 
## , , island = Torgersen
## 
##       species
## year   Adelie Chinstrap Gentoo
##   2007     10         0      0
##   2008     12         0      0
##   2009      8         0      0
```

It is observed that Adelie species can be found on every island, Gentoo species are only found on Biscoe island and Chinstrap could be found on Dream island.
The number of penguins on Biscoe island have increased over the years.

2. Estimating probability/proportions for Penguin population.

```
par(mfrow = c(1, 2))
hist = hist(my.penguins$bill_depth_mm[my.penguins$island=='Dream'],
            breaks = 20, freq = FALSE, xlab = "Bill Depth of Dream island penguins", main = "Histogram"
```



We find Mean and Standard deviation using inbuilt function for plotting Normal distribution.

```
Dream_bill_depth <- my.penguins$bill_depth_mm[my.penguins$island == 'Dream']

mean_Dream <- mean(Dream_bill_depth)
sd_Dream <- sd(Dream_bill_depth)

mean_Dream
```

```
## [1] 18.29867
```

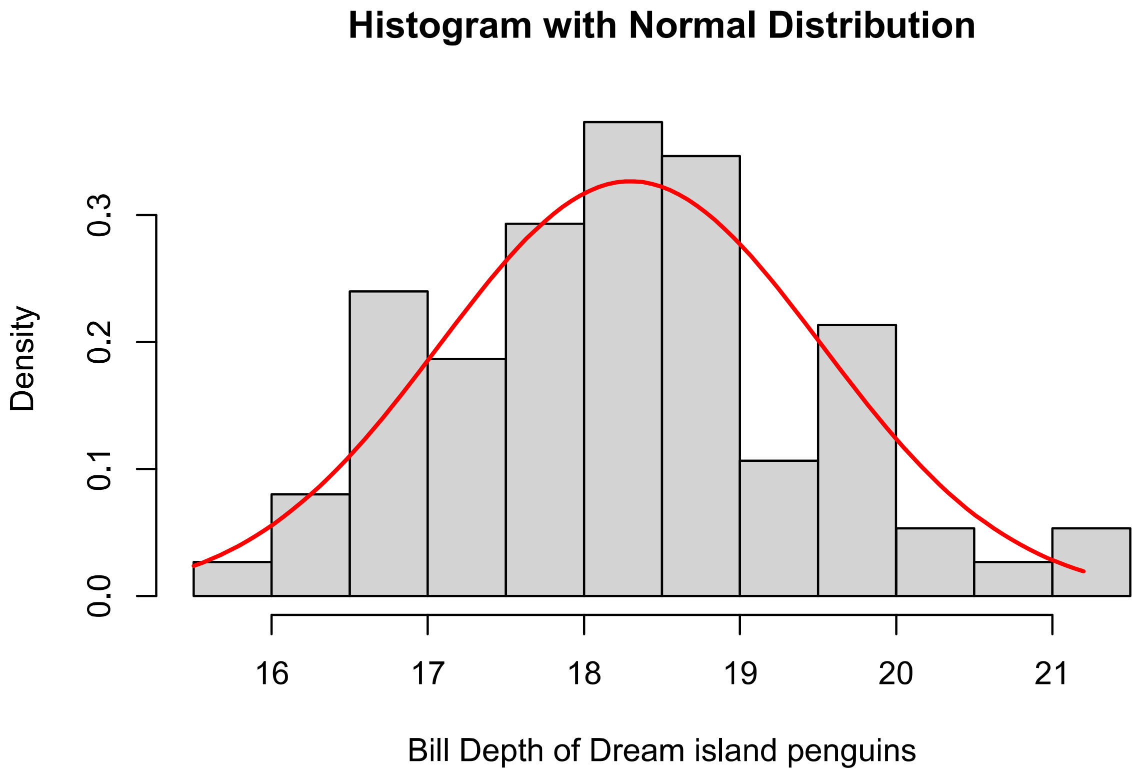
```
sd_Dream
```

```
## [1] 1.220932
```

```
# Plot histogram
hist(Dream_bill_depth, breaks = 20, freq = FALSE,
     xlab = "Bill Depth of Dream island penguins",
     main = "Histogram with Normal Distribution")

# values for the normal distribution curve
x <- seq(min(Dream_bill_depth), max(Dream_bill_depth), length = 100)
y <- dnorm(x, mean = mean_Dream, sd = sd_Dream)

# Overlay the normal distribution curve
lines(x, y, col = "red", lwd = 2)
```



The histogram reveals a near-normal distribution of bill depth for Dream Island's penguins. The computed mean and Standard deviation of this bill depth is found out to be 18.29867 and 1.220932 respectively. Following the construction of a normal distribution graph, it becomes evident that the body mass of penguins residing on Dream Island conforms to a normal distribution.

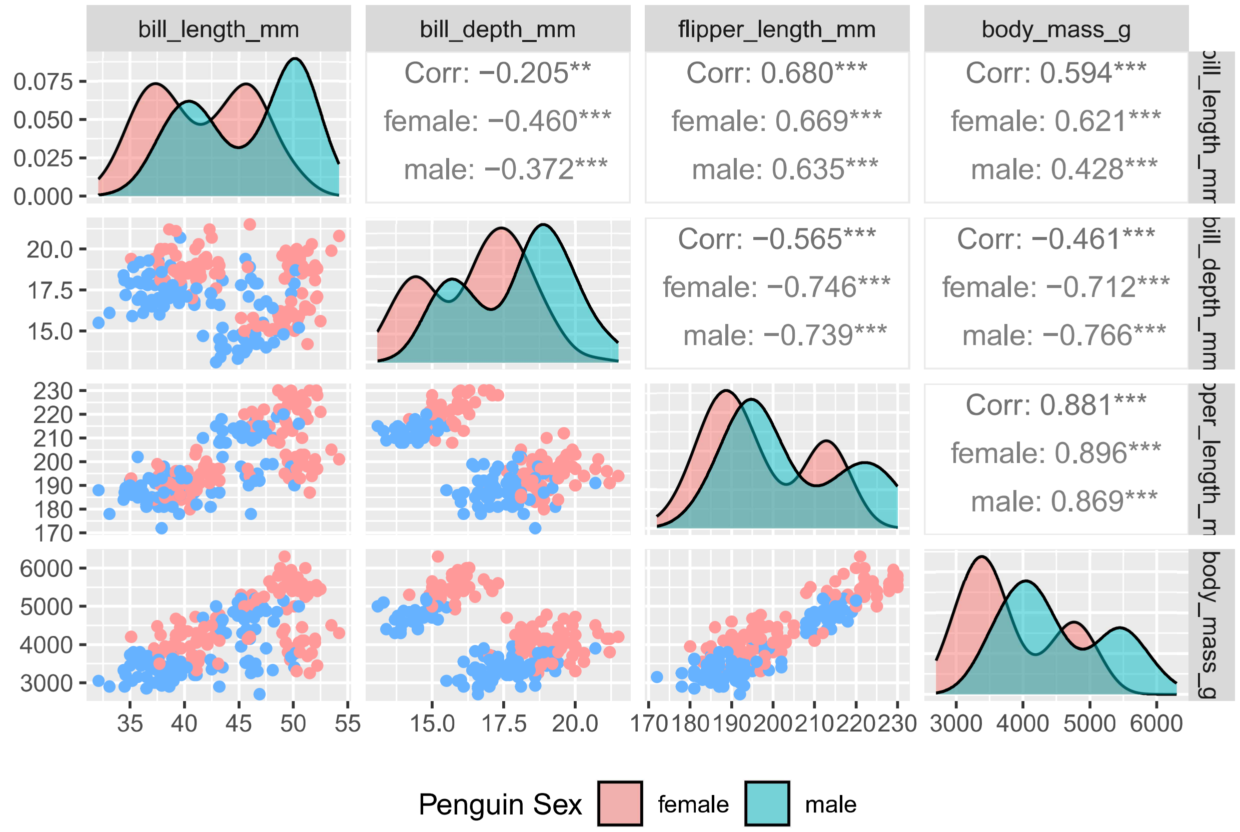
Usage of Maximum Likelihood Function to find mean and Standard deviation along with a Q-Q Plot could have been the best practice according to my research but due to my restricted knowledge in R-Programming I have taken the direct functions to find mean and Standard Deviation.

Maximum likelihood estimator(MLE) is considered best as it leverages the available data optimally to estimate the model parameters. This makes it a preferred choice for a wide range of applications, even in situations where assumptions of other models are not met.

A major disadvantage of MLE is that it tends to be biased for small samples of data.
To find the parameters of other measurement variables we need to first convert the data to a normal distribution by applying appropriate transformations.

3. Estimation of sex of a penguin from measurement data.

```r
ggpairs(my.penguins, columns = c(3:6), aes(color = sex), legend = 1,
        diag = list(continuous = wrap("densityDiag", alpha = 0.5 )), progress = FALSE) +
  theme(legend.position = "bottom") + labs(fill = "Penguin Sex") + scale_color_manual(values = c("male"
```



The graph shows relationships between four key physical measurements of penguins (bill length, bill depth, flipper length, and body mass), differentiated by sex. From the above plot we can observe that male penguins have longer flippers and higher body mass.
The data distribution of all continuous variables across genders is asymmetric and the data is multi-modal.

1. The scatter plots show positive correlations between most measurements, particularly strong between flipper length and body mass. This suggests that larger penguins tend to have larger measurements across all features.
2. There's a clear separation between males (pink) and females (blue) in most plots, indicating sexual dimorphism in penguin morphology. Males generally have larger measurements across all variables.
3. The diagonal density plots show the distribution of each measurement by sex. Most exhibit bimodal distributions, further emphasizing the sexual differences.
4. Bill length and depth show interesting patterns, with some overlap between sexes but still has differences.
5. Some plots reveal potential outliers, which could represent measurement errors or maybe slight data error.

Hypothesis test - Two Sample t-test

```
t.test(bill_length_mm ~ sex, data = my.penguins)
```

```
##
##  Welch Two Sample t-test
##
## data:  bill_length_mm by sex
## t = -5.9864, df = 197.94, p-value = 9.923e-09
## alternative hypothesis: true difference in means between group female and group male is not equal to
## 95 percent confidence interval:
##  -5.556903 -2.803021
## sample estimates:
## mean in group female   mean in group male
##             41.57732             45.75728
```

Test Statistic: t = -5.9864 This is the calculated t-value, which measures the difference between the two groups relative to the variation in the data.

Degrees of Freedom: df = 197.94 This indicates the sample size and variability in the data.

P-value: p-value = 9.923e-09 This is an extremely small p-value (much less than 0.05), indicating strong statistical significance.

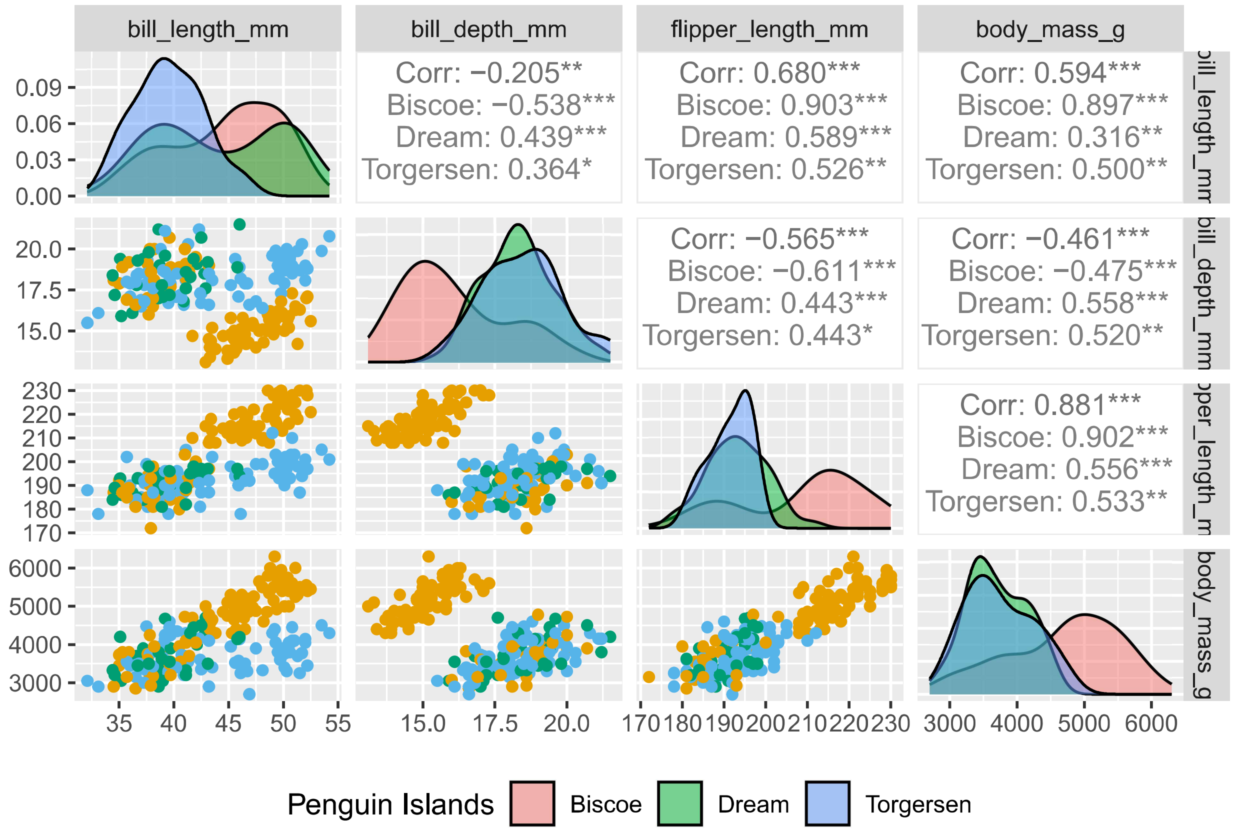
Confidence Interval: [-5.556903, -2.803021] This range represents the 95% confidence interval for the true difference in mean bill length between females and males.

Sample Estimates: Mean bill length for females: 41.57732 mm Mean bill length for males: 45.75728 mm

The very low p-value (9.923e-09) indicates that the difference in bill length between male and female penguins is highly statistically significant. Males have longer bills than females. We can be 95% confident that the true population difference in mean bill length between females and males falls between 2.80 and 5.56 mm less for females. Finally this test concludes that MALES HAVE SIGNIFICANTLY LONGER BILLS THAN FEMALES.

4. Significant difference in the physical characteristics of penguins living on different islands

```r
ggpairs(my.penguins, columns = c(3:6), aes(color = island), legend = 1,
        diag = list(continuous = wrap("densityDiag", alpha = 0.5 )),
        progress = FALSE) + theme(legend.position = "bottom") +
  labs(fill = "Penguin Islands") + scale_color_manual(values = c("Biscoe" = "#E69F00","Dream" = "#56B4E
```



Using this plot we can visually explore the correlations and distributions of all the continuous variables for penguins living on different islands. Body mass of penguins from dream island seem to have a normally distributed data. Penguins living on Biscoe island have longer flipper length and high body mass but lowest bill depth. Bill length and depth have gotten a weak to moderate correlation. Body mass has a strong positive correlation with flipper length but weak correlation with bill depth.

1. The diagonal density plots show distinct distributions across the three islands and also infers that penguin morphology varies significantly between islands.
2. Penguins from Biscoe Island (orange) often show a bimodal distribution in measurements, particularly in flipper length and body mass.
3. The scatter plots reveal positive correlations between flipper length and body mass. This correlation appears consistent across all islands.
4. In several plots, especially those involving flipper length or body mass, there's clear segregation between islands. Biscoe Island penguins tend to have larger measurements, while Torgersen Island penguins are often smaller.
5. Bill length and depth show interesting patterns, with some overlap between islands but has differences. Dream Island penguins seem to have a wider range of bill depths compared to the other islands.

Hypothesis test - ANOVA (Analysis of Variance) test

```r
aov(bill_length_mm ~ island, data = my.penguins)
```

```
## Call:
##    aov(formula = bill_length_mm ~ island, data = my.penguins)
##
## Terms:
##                   island Residuals
## Sum of Squares   638.728  5079.132
## Deg. of Freedom        2       197
##
## Residual standard error: 5.077637
## Estimated effects may be unbalanced
```

Model Formula: bill_length_mm ~ island This indicates that we're analyzing how bill length varies based on the island where the penguins are found.

Island: 638.728 Residuals: 5079.132 The Sum of Squares for 'island' represents the variation in bill length that can be explained by differences between islands.

Degrees of Freedom: Island: 2 (This is because there are three islands, so 3 - 1 = 2) Residuals: 197

Residual standard error: 5.077637 This is an estimate of the standard deviation of the residuals.

ANOVA provides evidence that BILL LENGTH VARIES SIGNIFICANTLY AMONG PENGUINS FROM DIFFERENT ISLANDS, but island location alone explains only a modest portion of the overall variation in bill length.

Conclusion

The dataset was successfully analyzed using a variety of statistical approaches. The examination of the exploratory data revealed important information on the connections between various factors. We were able to calculate the mean and standard deviation. Non-parametric tests were conducted to investigate how different variables relate to the sex and island of the penguins.

The results indicate that certain physical characteristics, including bill length, bill depth, flipper length, and body mass, exhibit a higher correlation with the sex of the penguins. This suggests that these variables can be utilized by scientists for sex determination in penguins. However, Welch Two Sample t-test test concludes that MALES HAVE SIGNIFICANTLY LONGER BILLS THAN FEMALES. Furthermore, it doesn't seem that the year of measurement is connected to the penguin's sex. Therefore, determining sex would not benefit from knowing the island, species, or year.

Based on the results of our exploratory data analysis and the ANOVA (Analysis of Variance) test, we may conclude that the physical attributes of Dream and Torgersen island penguins are different. ANOVA provides evidence that BILL LENGTH VARIES SIGNIFICANTLY AMONG PENGUINS FROM DIFFERENT ISLANDS.