Statistics-PalmerPenguins

Varsha Venkataraman - 240150459

2024-10-24

The dataset that is used in this analysis comprises **Palmer penguins** which comprises measurement from a subset of **200 penguins** that is sampled from a population of three species: **Chinstrap, Gentoo, and Adelie**. The dataset also included details of the penguins' physical features: **bill length, bill depth, flipper length, and body mass**. Moreover, the dataset records the islands of origin(**Dream, Biscoe, or Torgerson**) and the gender(**male,female**) of each penguin. This analysis aims to achieve all the task that are mentioned in the coursework.

Table 1: Summary Statistics of Penguin Features

Statistic	Body.Mass	Bill.Depth	Bill.Length	Flipper.Length
Range	3200.00	8.40	27.50	59.00
Variance	616918.95	3.98	31.63	195.36
Mean	4193.88	17.14	43.99	43.99
Standard Deviation	785.44	1.99	5.62	13.98
IQR	1125.00	3.05	9.02	23.25
Coefficient of Variation	18.73	11.64	12.78	6.95

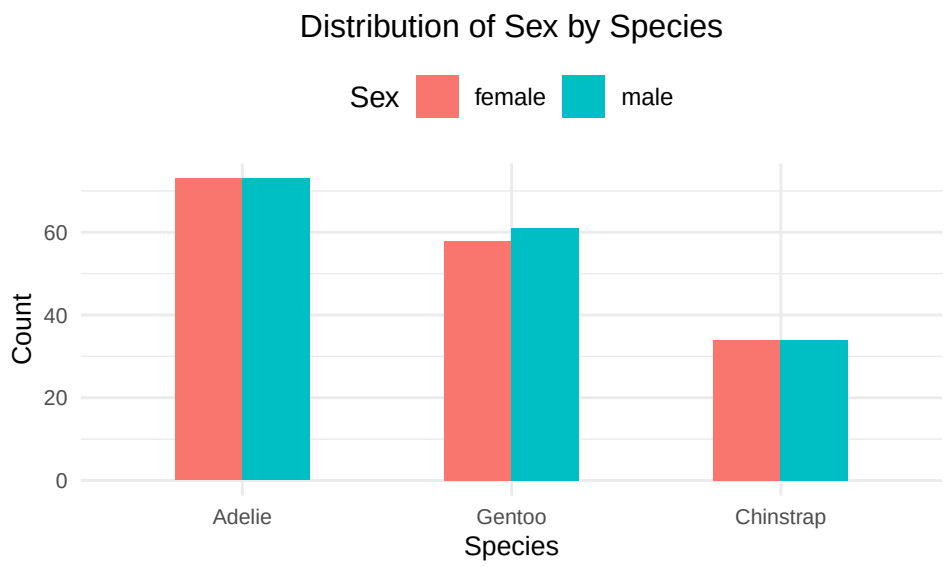
Table 2: Summary Statistics of Penguin Features

Statistic	Bill_Length	Bill_Depth	Flipper_Length	Body_Mass
Min	32.10	13.10	172.00	2850.0
Q1	39.50	15.65	190.00	3587.5
Median	44.55	17.25	197.00	4050.0
Q3	48.52	18.70	213.25	4712.5
Max	59.60	21.50	231.00	6050.0

Task 01: Summarise the Physical Attributes of Penguins

Numerical Analysis - Results from our numerical analysis of the Palmer penguin's physical characteristics reveal that many aspects, especially body mass (**Tables 1 and 2**), show very large range and high standard deviation which indicates unique variability in traits. So it is summarized that the environmental factors probably denoted by diet and habitat may affect body size of penguin

In contrast, bill depth exhibits little variation and reflects a common diet-related adaptation among individuals. The varying size of penguins' flippers suggest the changes are related to modifications for swimming economy. A wider range of bill length also means potential environmental pressures related to different feeding behavior. From the unique coefficients of variation it can be attributed that they exhibit different ranges of standard variance within the population.



Graphical Analysis: Barplot showing the number of female and male penguins per species: Adelie, Chinstrap or Gentoo. From the plot, it can be seen that similar number of male and female penguins are there in Adelie as well as Chinstrap species. The Gentoo species however displayed an odd quirk in their numbers; more male than female penguins. This disparity in the Gentoo race will be a clear indication that there must be gender imbalance.

I ploted this barplot for **Task 4**, as Adelie and Chinstrap almost have equal gender split, so I assume that there is unique difference in the physical characteristics between the genders.

Task 02: Fitting a Probability Distribution to Penguin Measurements

Box-Cox Transformation and Log-Likelihood Calculation

The Box-Cox transformation stabilizes variance and makes data more normal: \section

$$Y' = \begin{cases} \frac{Y^\lambda - 1}{\lambda} & \lambda \neq 0, \\ \log(Y) & \lambda = 0. \end{cases}$$

The optimal λ is obtained by maximizing the log-likelihood:

$$\lambda_{\text{opt}} = \arg\max_{\lambda} \ \log \mathcal{L}(\lambda | Y).$$

For our dataset: - Optimal λ: **1.515152** - Mean (μ): **0.0032** - Variance (σ^2): **0.9533**

The optimal Box-Cox transformation for bill depth yielded $\lambda = 1.515152$, indicating a need to stabilize variance.

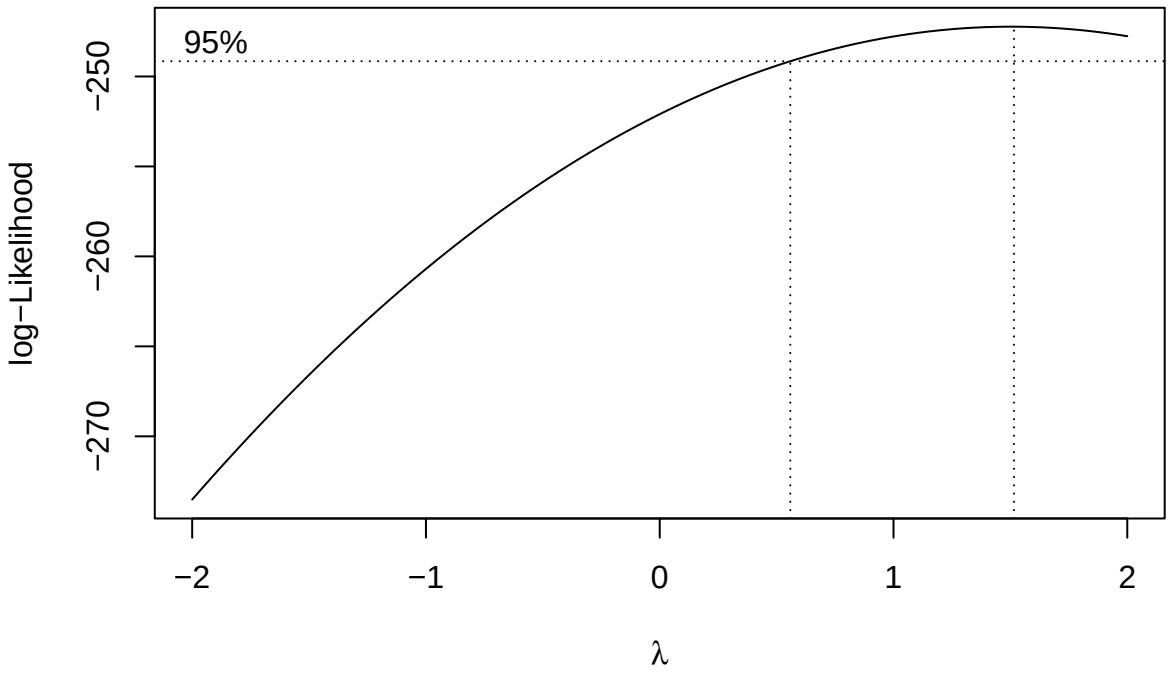
Normal Distribution Fit for Box-Cox Transformed Bill Depth

The probability density function (PDF) of a normal distribution is defined as:

$$f(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

where: x represents the Box-Cox transformed bill depth measurements.

- μ is the mean of the transformed data, calculated as: $\mu = \frac{\sum_{i=1}^{n} x_i}{n} = 48.59377$.
- σ is the standard deviation of the transformed data, calculated as: $\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}} = 8.474398$.

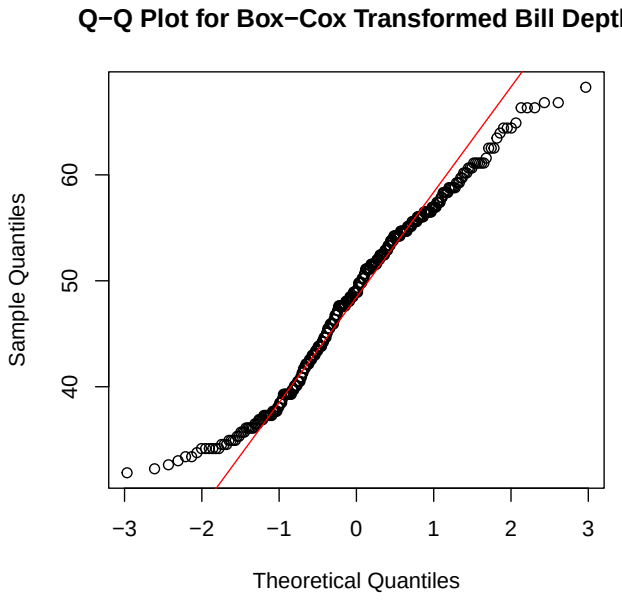


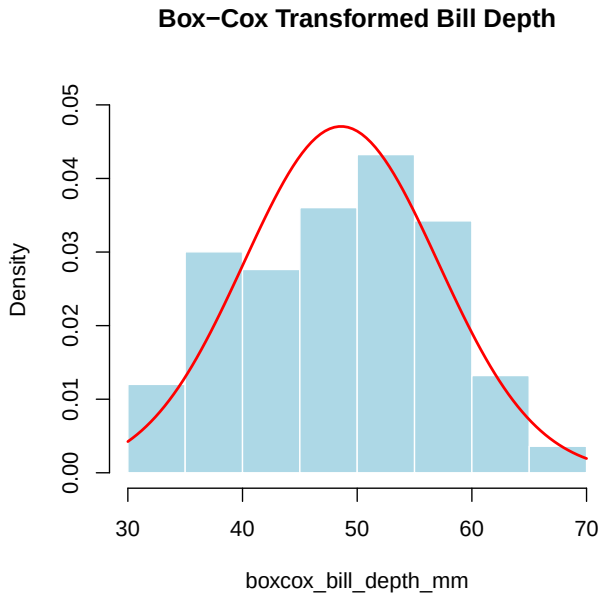

```r
lambda_opt <- boxcox_result_bill_depth$x[which.max(boxcox_result_bill_depth$y)]
boxcox_transformed <- (penguins$bill_depth_mm^lambda_opt - 1) / lambda_opt
mean_transformed <- mean(boxcox_transformed)
sd_transformed <- sd(boxcox_transformed)
```

Using these values, the normal curve is fitted over the histogram of the Box-Cox transformed bill depth. This provides a visual check of how well the transformed data aligns with a normal distribution.

For our dataset: $\mu = 48.6$ and $\sigma = 8.5$

Metric	Value
Shapiro-Wilk Test p-value	< 0.001





However, the Shapiro-Wilk test returned a p-value < 0.001, suggesting that even the transformed data is not normally distributed. This result is further supported by the Q-Q plot, which indicated deviations from normality. Given these results, using the normal distribution to estimate probabilities for the penguin population may not provide accurate estimates. **Instead, alternative distributions like the log-normal or gamma could better capture the characteristics of the data, especially given the observed skewness**.

Task 03: Key Variables for Distinguishing Penguin Sex

Table 4: Bartlett's Test Results for Homogeneity of Variances by Island

Variable	K_squared	p_value
body_mass_g	60.99966	< 0.001
bill_depth_mm	30.60355	< 0.001
bill_length_mm	25.36815	< 0.001
flipper_length_mm	76.39295	< 0.001

Table 5: Bartlett's Test Results for Homogeneity of Variances by Sex

Variable	K_squared	p_value
body_mass_g	4.6016841	0.0319
bill_depth_mm	0.2256684	0.6348
bill_length_mm	1.3429293	0.2465
flipper_length_mm	3.7766600	0.0520

Table 6: T-Test Results for Sex Distinction

Comparison	T.Statistic	P.Value
body_mass_g by Sex (Male vs Female)	-8.5417	< 0.001
bill_depth_mm by Sex (Male vs Female)	-7.3065	< 0.001
bill_length_mm by Sex (Male vs Female)	-6.6670	< 0.001
flipper_length_mm by Sex (Male vs Female)	-4.8013	< 0.001

Bartlett's Test : This test is performed here to check if the variances of variables (e.g., body mass, bill depth) are similar across different groups (e.g., by island or by sex).

Two groups mean t-tests – It is being used to determine if the means of different physical traits (body mass and bill depth for male vs female penguins) are significantly equal contrarily.

A p-value below 0.05 in a t-test is indicative of the two group's means being significantly different from each other, and it tells us that such an observation would have not come randomly into existence given our null hypotheses consisting on both data sets coming from the same source.

Table 7: Logistic Regression Model Summary

term	estimate	std.error	statistic	p.value
(Intercept)	-59.0511378	9.0527392	-6.523013	< 0.001
body_mass_g	0.0056144	0.0008920	6.294457	< 0.001
bill_depth_mm	2.0670075	0.3228626	6.402127	< 0.001

Logistic regression is used when the dependent variable is binary (in this case, "sex" which can be "male" or "female"). It models the probability of a certain class (e.g., being "male") given the values of predictor variables (e.g., body mass and bill depth). It allows for estimating the odds of a penguin's sex based on its physical measurements.This helps in understanding how the characteristics like body mass and bill depth influence the likelihood of a penguin being classified as male or female.

Table 8: Logistic Regression Model Summary - Accuracy Summary

Metric	Value
Model Accuracy	87.5%
Correct Predictions	175
Incorrect Predictions	25

With respect to **(Table 8)** we can conclude that the logistic regressionmodel achieves an accuracy of **87.5%** which classifies the sex based on the key variables bill depth and body mass.

Task 04: Impact of Island Origin on Penguin Physical Characteristics

With respect to **(Table 4)** the results suggest that the variance of these physical traits is not consistent across different islands. This variance in physical characteristics supports the findings from other tests, such as **the t-test and MANOVA**, which indicate that the island of origin plays a significant role in the physical differences observed in penguins

According to **(Table 9)** the t-test results suggest that the island of origin has a significant impact on some physical characteristics of penguins, particularly when comparing Torgersen and Biscoe or Biscoe and Dream. However, the differences between Torgersen and Dream are not significant for most of the traits, indicating that the island's impact varies depending on the specific characteristic and island pair being compared.

After t-test, proceeding with MANOVA gives a broader understanding of whether the combination of traits varies across islands as a whole. This allows for a more complete analysis of how the island of origin impacts the physical characteristics of penguins.

MANOVA analysis, According to **(Table 10)** Wilks' Lambda was used to evaluate whether there are significant differences in physical characteristics of penguins (bill depth, bill length, flipper length, and body mass) based on their island of origin. A low Wilks' Lambda value, along with a significant p-value, suggests that the mean vectors of these characteristics differ significantly across islands. This indicates that the physical traits of penguins are influenced by the island they inhabit.

Table 9: T-Test Results Between Islands

Comparison	T.Statistic	Degrees.of.Freedom	P.Value
Torgersen vs Biscoe	8.3669	208	< 0.001
Torgersen vs Dream	0.1429	168	0.8865
Biscoe vs Dream	12.7712	284	< 0.001
Torgersen vs Biscoe	-8.8666	208	< 0.001
Torgersen vs Dream	-0.5415	168	0.5889
Biscoe vs Dream	-12.9841	284	< 0.001
Torgersen vs Biscoe	8.3504	208	< 0.001
Torgersen vs Dream	5.6929	168	< 0.001
Biscoe vs Dream	1.6103	284	0.1084
Torgersen vs Biscoe	8.4144	208	< 0.001
Torgersen vs Dream	1.3559	168	0.177
Biscoe vs Dream	11.5819	284	< 0.001

Table 10: MANOVA Test Results

Test	approx_F	df	p_value
Wilks' Lambda	51.5022	8	0

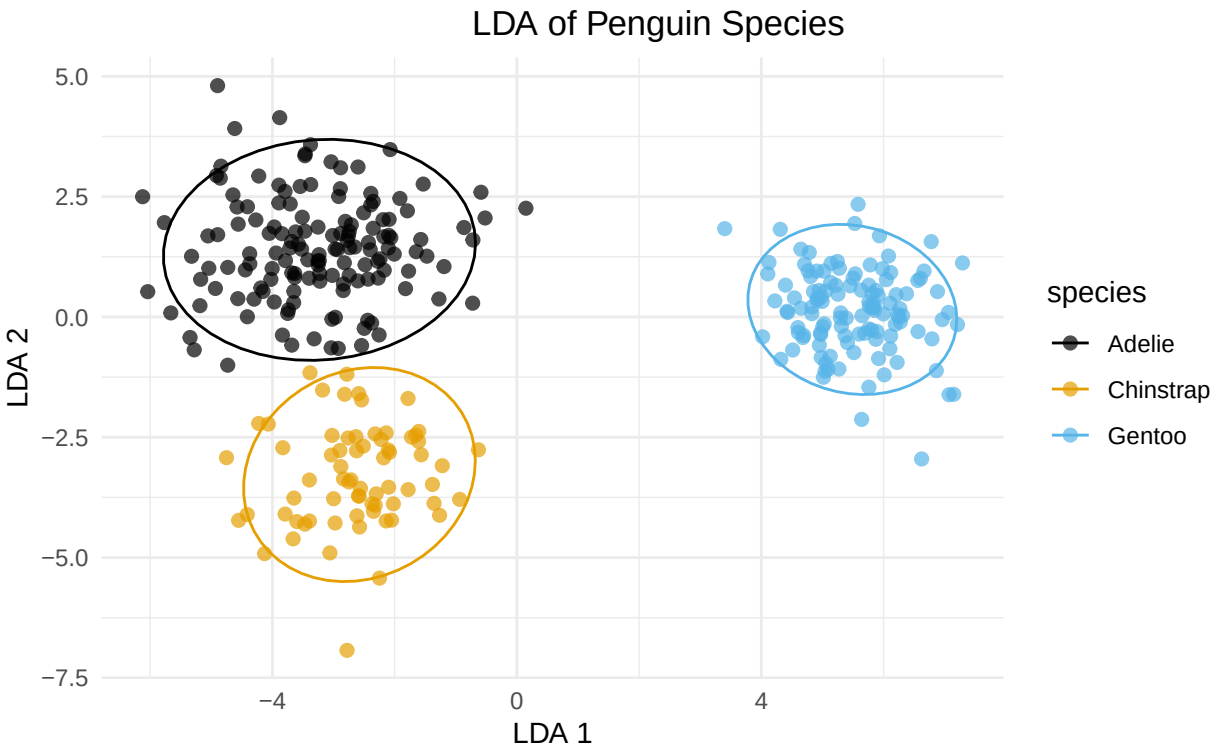
The results from other MANOVA tests, such as **Pillai's Trace, Hotelling-Lawley Trace, and Roy's Largest Root**, supported similar conclusions, affirming the significant impact of the island on the combined physical traits. Each test provided evidence that variations in body mass, flipper length, and bill measurements are associated with differences in the island environment, which could relate to ecological adaptations or evolutionary factors. The **correlation analysis** is conducted after MANOVA to explore the interrelationships between different physical characteristics of penguins.

Overall, as indicated by **(Table 11)**, the findings suggest that while some physical characteristics (e.g., body mass and flipper length) are positively correlated, bill length negatively correlates with bill depth, indicating that as one increases, the other tends to decrease. The significant correlation supports the hypothesis that the physical characteristics of penguins are interrelated, which is essential for understanding their variation across different islands.

Table 11: Correlation Summary of Penguin Features

Feature_1	Feature_2	Correlation_Coefficient
bill_depth_mm	bill_length_mm	-0.23
bill_depth_mm	flipper_length_mm	-0.58
bill_depth_mm	body_mass_g	-0.47
bill_length_mm	bill_depth_mm	-0.23
bill_length_mm	flipper_length_mm	0.65
bill_length_mm	body_mass_g	0.59
body_mass_g	bill_length_mm	0.59
body_mass_g	bill_depth_mm	-0.47
body_mass_g	flipper_length_mm	0.87
flipper_length_mm	bill_length_mm	0.65
flipper_length_mm	bill_depth_mm	-0.58
flipper_length_mm	body_mass_g	0.87

Linear Discriminant Analysis (LDA) is a statistical classification technique used to analyze and classify data based on a set of features. Feature Selction ,with respect to **(Table 11)** the flipper length and bodymass are strong features for inclusion in LDA model due to their strong positive correlation. I am considering bill length , flipper length as the Variation Inflation Factor calculation is 1.74 which is not problamatic so it is retained for consideration, body mass and bill_depth_mm are as well considered.



The LDA model **(Table 12)** achieved a perfect classification accuracy of 100%, indicating that it can effectively **distinguish between the different penguin species based on the selected features**. The centroid values indicate that the LDA model provides a clear separation between Gentoo and the other two species primarily through LD1, while LD2 further separates Adelie from Chinstrap.

This analysis provides a comprehensive view of how well the model distinguishes between different penguin species based on the selected features.

Table 12: Summary of LDA Model Accuracy and Centroids

Species	Centroid_LD1	Centroid_LD2	Model_Accuracy
Adelie	-3.267005	1.4229569	100%
Chinstrap	-2.632881	-3.2930120	100%
Gentoo	5.512761	0.1359085	100%

Conclusion:

The **Linear Discriminant Analysis** successfully classified penguin species based on physical characteristics, achieving high accuracy and revealing distinct centroids for each species in the LDA space. This indicates a strong relationship between the selected features and species classification, supporting the research question

References:

An introduction to applied multivariate statistics , Muni Shanker Srivastava 1936-. E. M Carter (Edward M.). c1983 - Available at **Philip Robinson library**

Quantitative social science data with R - Brian J. Fogarty - Available at **Philip Robinson library**

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer.