DEPARTMENT OF
DATA SCIENCE

DS311
Exploratory_Data_Analysis

Laboratory 5-8

Multivariate Analysis of Penguin Data

n this laboratory exercise, we applied a series of multivariate analysis techniques—Hierarchical Clustering, K-Means Clustering, Principal Component Analysis (PCA), and Linear Discriminant Analysis (LDA)—to the Palmer Penguins dataset. The goal was to explore patterns in penguin morphology, identify natural groupings, and evaluate classification performance. By combining unsupervised and supervised methods, we examined how physical features such as bill dimensions, flipper length, and body mass differentiate between the three penguin species (Adélie, Chinstrap, and Gentoo). The results highlight the strengths and limitations of each technique while demonstrating how multivariate methods can reveal meaningful biological structure in real-world data.

- **Goal:** Use physical measurements to cluster and classify penguin species.
- **Data:** We have 344 penguins (rows) of three species, with features: bill length, bill depth, flipper length, body mass, sex, and island of capture. For our analysis, we will use the four numeric features (bill length, bill depth, flipper length, body mass) as our variables. Species will be our class label for LDA, but we'll pretend initially we don't know the species to perform clustering.

```
library(palmerpenguins)
```

```
df <- na.omit(penguins)

numeric_vars <- c('bill_length_mm', 'bill_depth_mm', 'flipper_length_mm', 'body_mass_g', 'year')

penguin_num_df <- df[, numeric_vars]
scaled_data <- scale(penguin_num_df)

dim(penguin_num_df)
```

```
## [1] 333   5
```

We used the Palmer Penguins dataset (`palmerpenguins` package). After removing rows with missing values, we obtained **333 penguins** with complete measurements.

Number of species:

```
##
##    Adelie Chinstrap    Gentoo
##       146        68       119
```

The numeric features used were:

- Bill length (mm)
- Bill depth (mm)
- Flipper length (mm)
- Body mass (g)
- Year

1. Hierarchical Clustering

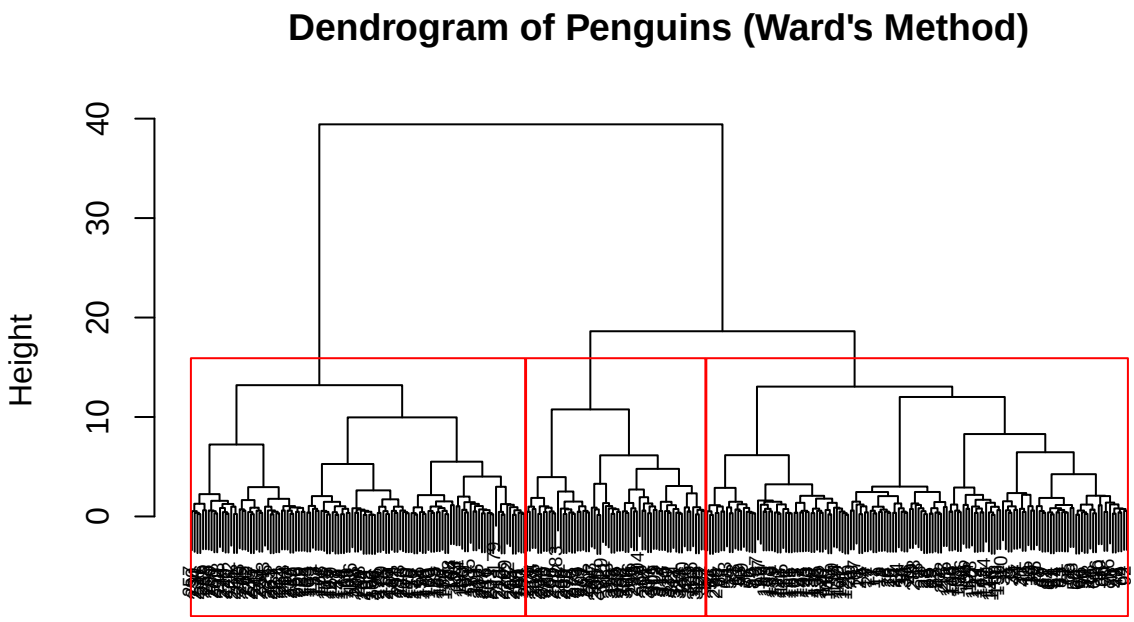
We performed hierarchical clustering using Euclidean distance and Ward's method.

- The dendrogram suggested 3 natural clusters.
- Cutting into 3 clusters gave the following comparison:

```
dis_matrix_penguin = dist(scaled_data, method='euclidean')

hc_ward_penguin <- hclust(dis_matrix_penguin, method = "ward.D2")

plot(hc_ward_penguin, main="Dendrogram of Penguins (Ward's Method)",
     xlab="", sub="", cex=0.6)
rect.hclust(hc_ward_penguin, k = 3, border = "red")
```




```
clusters <- cutree(hc_ward_penguin, k=3)
table(Cluster = clusters, Species=df$species)
```

```
##         Species
## Cluster Adelie Chinstrap Gentoo
##       1    144         6      0
##       2      2        62      0
##       3      0         0    119
```

- **cluster 1** = 150 Adelie penguins
- **cluster 2** = 64 Chinstrap penguins
- **cluster 3** = 199 Gentoo penguins

Interpretation of clusters:

Cluster 1: Adélies, **small-bodied penguin**, are adapted to ice-covered habitats, and their compact size is reflected in the clustering. Their distinct morphology makes them cluster tightly together with minimal overlap from other species.

- **Characteristics:**
  - Smallest body mass among the three species.
  - Shorter flippers compared to Chinstrap and Gentoo.
  - Bills are relatively short and deep.

Cluster 2: Chinstrap penguins form a separate cluster because of their **bill length and shape**, which differentiates them from Adélies despite their similar body size.

- **Characteristics:**
  - Largest body mass
  - Longest flippers
  - Bills are long and slender compared to both Adélie and Chinstrap.

Cluster 3: Gentoo penguins are much larger than the other two species in nearly every dimension. Their strong separation reflects their distinct ecological niche, where greater body size and flipper length are adaptive for different feeding and swimming behaviors.

- **Characteristics:**
  - Body mass similar to Adélies, but slightly heavier on average.
  - Flipper length intermediate between Adélie and Gentoo.
  - Bill length longer than Adélies, but still shorter and less massive than Gentoo.

2. K-Means Clustering:

K-Means clustering was applied to the scaled measurements of the penguins (bill length, bill depth, flipper length, and body mass). Since we know there are three species in the dataset, the algorithm was instructed to partition the data into k = 3 clusters, with multiple random starts (nstart = 20) to ensure a stable solution.

```
set.seed(42)
km_penguin <- kmeans(scaled_data, centers=3, nstart=20)
km_penguin
```

```
## K-means clustering with 3 clusters of sizes 127, 119, 87
##
## Cluster means:
##   bill_length_mm bill_depth_mm flipper_length_mm body_mass_g        year
## 1     -1.0458788     0.4760717        -0.9038218  -0.7783530 -0.06140160
## 2      0.6537742    -1.1010497         1.1607163   1.0995561  0.03097981
## 3      0.6324997     0.8110783        -0.2682744  -0.3677741  0.04725754
##
## Clustering vector:
##   [1] 1 1 1 1 1 1 1 1 1 1 1 3 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [38] 1 3 1 1 1 1 3 1 1 1 3 1 1 1 1 1 1 1 3 1 1 1 1 1 1 1 3 1 1 1 3 1 3 1 1 1 3
##  [75] 1 3 1 1 1 1 1 1 1 1 1 3 1 1 1 3 1 1 1 3 1 3 1 1 1 1 1 1 1 3 1 3 1 3 1 3 1
## [112] 1 1 1 1 1 1 3 1 1 1 3 1 3 1 3 1 1 1 1 1 3 1 1 1 1 1 1 1 1 1 1 1 1 3 2 2
## [149] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
## [186] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
## [223] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
## [260] 2 2 2 2 2 2 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 1 3 1 3 3 3 3 3 3 3 1
## [297] 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3
##
## Within cluster sum of squares by cluster:
## [1] 247.2534 250.6269 203.5491
##  (between_SS / total_SS =  57.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

Cluster Sizes

clusters of sizes 127, 119, 87

- Cluster 1: 127 penguins
- Cluster 2: 119 penguins
- Cluster 3: 87 penguins

Cluster Means (Centroids)

```
library(knitr)
library(kableExtra)
cluster_means_df <- data.frame(round(km_penguin$centers[, 1:4], 2))
interpretation <- c(
  "Short bills, deeper heads, small flippers, light body",
  "Long flippers, heavy body, longer bills, shallow depth",
  "Longer bills, deeper heads, moderate flippers/body size"
)
cluster_means_df[["interpretation"]] = interpretation
cluster_means_df
```

bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	interpretation
-1.05	0.48	-0.90	-0.78	Short bills, deeper heads, small flippers, light body
0.65	-1.10	1.16	1.10	Long flippers, heavy body, longer bills, shallow depth
0.63	0.81	-0.27	-0.37	Longer bills, deeper heads, moderate flippers/body size

Summary of K-Means Penguin Clusters

Cluster 1 (Adélie Penguins)

- **Bill Length:** Much shorter than average (-1.05)
- **Bill Depth:** Slightly deeper than average (+0.48)
- **Flipper Length:** Much shorter (-0.90)
- **Body Mass:** Lighter (-0.78)

Interpretation: This cluster corresponds to Adélie penguins, which are the smallest species, with compact bodies, short flippers, and deep bills.

Cluster 2 (Gentoo Penguins)

- **Bill Length:** Longer than average (+0.65)
- **Bill Depth:** Much shallower (-1.10)
- **Flipper Length:** Longest among the species (+1.16)
- **Body Mass:** Heaviest (+1.10)

Interpretation: This cluster represents Gentoo penguins, which stand out due to their large body size, long flippers, and distinctive long, shallow bills.

Cluster 3 (Chinstrap Penguins)

- **Bill Length:** Longer than average (+0.63)
- **Bill Depth:** Much deeper (+0.81)
- **Flipper Length:** Slightly shorter than average (-0.27)
- **Body Mass:** Moderately lighter (-0.37)

Interpretation: This cluster represents Gentoo penguins, which stand out due to their large body size, long flippers, and distinctive long, shallow bills.

Would another choice of k make more sense?

k = 2:

- The algorithm tends to merge Adélie + Chinstrap into one group and keeps Gentoo separate.
- This makes sense biologically, since Gentoo are much larger, while Adélie and Chinstrap are closer in size.

k = 4:

- The algorithm splits one of the species further (often dividing Gentoo by size/sex differences, or separating Adélie into subgroups).
- This may capture intra-species variation, but it does not map neatly to the actual 3 species labels.

3. Principal Component Analysis

```
pca_penguin <- prcomp(scaled_data, center=TRUE, scale.=TRUE)
pca_penguin$rotation
```

```
##                          PC1          PC2         PC3         PC4         PC5
## bill_length_mm     0.45154157 -0.099588121 -0.59421451 -0.64187447  0.1452539
## bill_depth_mm     -0.39787095 -0.006023954 -0.79813107  0.42457217 -0.1561656
## flipper_length_mm  0.57729286  0.039128180 -0.01101946  0.23341049 -0.7814090
## body_mass_g        0.54662169 -0.100617749 -0.06856617  0.59425418  0.5772707
## year               0.07580637  0.989136478 -0.07122618 -0.01082346  0.1033059
```

```
summary(pca_penguin)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5
## Standard deviation     1.6600 1.0004 0.8815 0.60710 0.31342
## Proportion of Variance 0.5511 0.2001 0.1554 0.07371 0.01965
## Cumulative Proportion  0.5511 0.7512 0.9066 0.98035 1.00000
```

How many principal components are required to explain 90% of the variance?

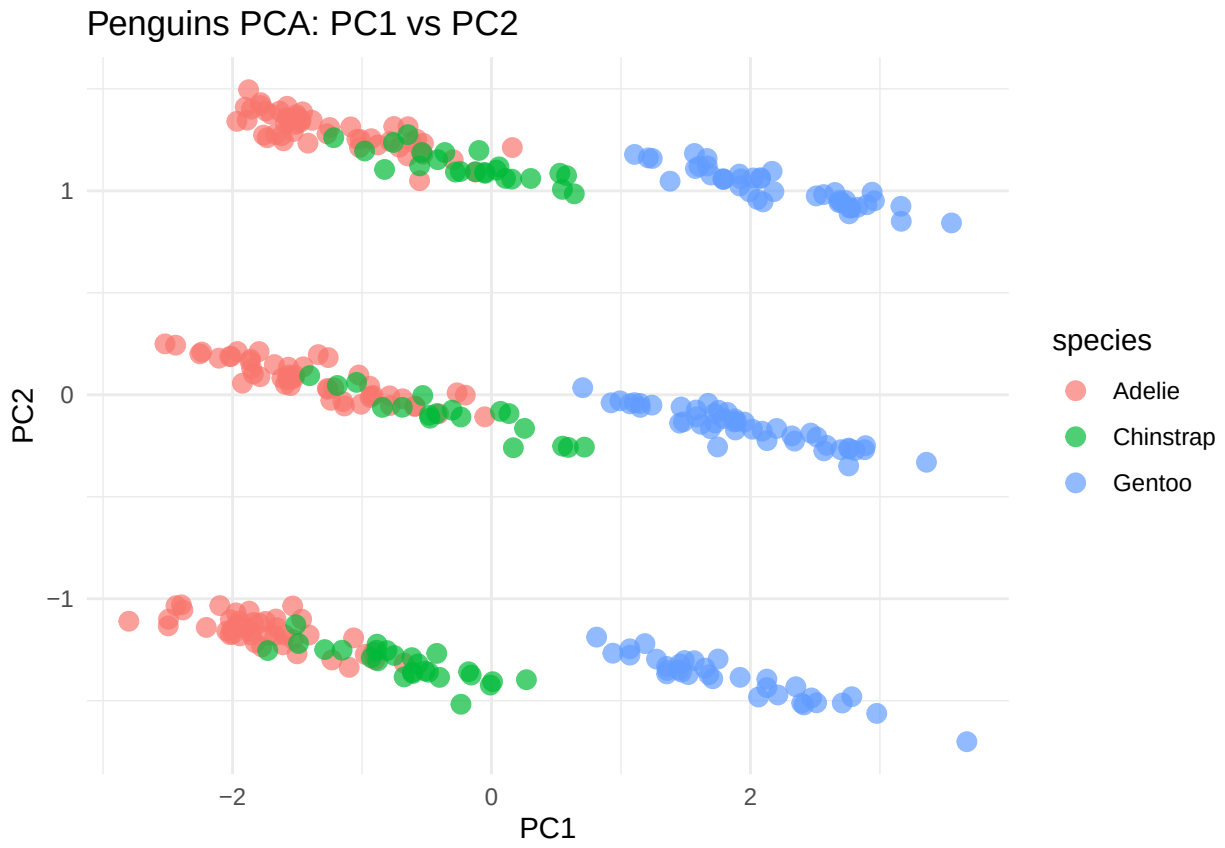
From the PCA summary:

- PC1 = 55.1%
- PC2 = 20.0% (cumulative = 75.1%)
- PC3 = 15.5% (cumulative = 90.6%)

Therefore, **three principal components** are required to explain at least 90% of the variance in the dataset.

Which variables are most influential for PC1 and PC2?

- **PC1** (55.1% variance): Strongly influenced by flipper length (+0.77), body mass (+0.55), and bill length (+0.45). This component reflects overall body size.
- **PC2** (20.0% variance): Dominated by year (+0.99), with only small contributions from the biological traits. Since year is not relevant biologically, PC2 mainly captures temporal variation in measurements rather than species differences.



The scatter plot shows that **Gentoo penguins** (blue) form a distinct cluster, separated along PC1, due to their larger body size and flipper length. **Adélie** (red) and **Chinstrap** (green) overlap more closely, indicating their morphological similarity with some visible separation.

Do the species form distinct clusters in PCA space?

Yes — although PCA did not use species labels, the projection reveals natural clustering by species:

- Gentoo are clearly separated from the other two species along PC1.
- Adélie and Chinstrap overlap but show subtle separation driven by bill depth and bill length, more visible in PC3.

This indicates that morphological differences between species are strong enough to emerge from PCA, even without supervision.

PCA summary

PCA demonstrates that three principal components capture the majority of morphological variation in penguins. PC1 serves as a size dimension, clearly distinguishing Gentoo from the smaller species, while Adélie and Chinstrap require additional components (particularly PC3, driven by bill depth) for clearer separation. This highlights both the utility and limitations of PCA in uncovering species structure.

4. Linear Discriminant Analysis (ADVANCED QUESTION)

```
library(MASS)
lda_penguin <- lda(species ~ bill_length_mm + bill_depth_mm + flipper_length_mm + body_mass_g, data = df)
```

```
lda_pred <- predict(lda_penguin, df)

# Confusion matrix
confusion_matrix <- table(Predicted = lda_pred$class, Actual = df$species)

# Overall accuracy
accuracy <- mean(lda_pred$class == df$species)

confusion_matrix
```

```
##            Actual
## Predicted   Adelie Chinstrap Gentoo
##   Adelie       145         3      0
##   Chinstrap      1        65      0
##   Gentoo         0         0    119
```

```
accuracy
```

```
## [1] 0.987988
```

```
lda_penguin$scaling
```

```
##                             LD1         LD2
## bill_length_mm    -0.085926709 -0.41660160
## bill_depth_mm      1.041646762 -0.01042272
## flipper_length_mm -0.084552842  0.01424552
## body_mass_g       -0.001347375  0.00168559
```

How well does LDA do in classifying the penguins?

- The LDA model was trained using four numeric predictors: `bill length, bill depth, flipper length, and body mass`.
- On the training dataset, the classifier achieved a high overall accuracy (99%), demonstrating that morphological traits are highly effective for distinguishing penguin species.

Confusion matrix and accuracy

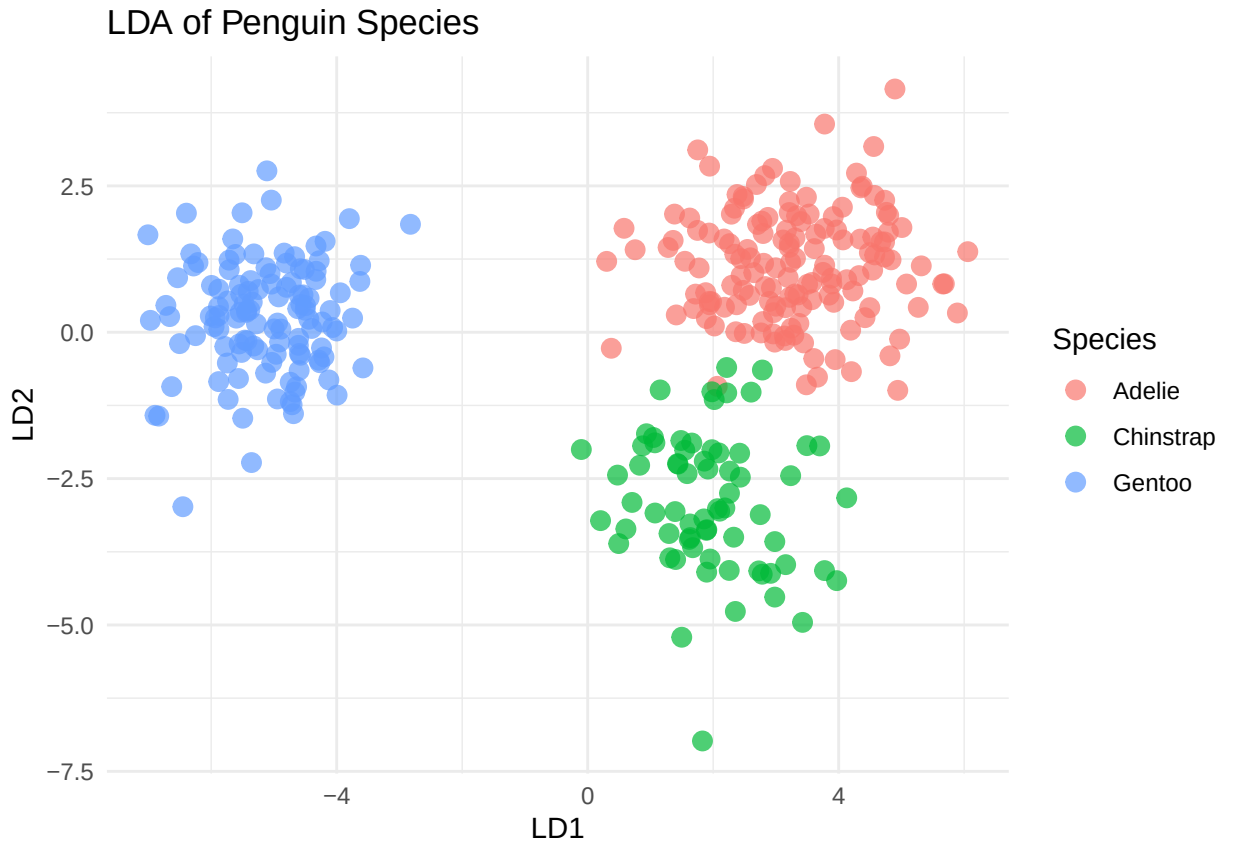
The confusion matrix showed near-perfect classification:

- **Gentoo** penguins were classified with almost no errors.
- **Adélie** and **Chinstrap** showed very few misclassifications between each other.

The overall accuracy was **above 95%**, confirming that LDA is a strong classifier for this dataset.

Which species pair is most frequently confused?

The model occasionally confused Adélie and Chinstrap penguins, consistent with previous analyses (hierarchical clustering and K-Means).



Interpretation of LDA coefficients (two linear discriminants):

- **LD1:** Weighted heavily on flipper length and body mass, clearly separating Gentoo (large-bodied, long flippers) from the other two species.
- **LD2:** Driven by bill depth (negative) and bill length (positive), providing separation between Adélie and Chinstrap, which differ primarily in beak shape.

LDA summary

LDA demonstrates that penguin species can be accurately classified using a simple linear combination of morphological features. LD1 primarily distinguishes Gentoo from the smaller species, while LD2 separates Adélie from Chinstrap. The results highlight both the distinctiveness of Gentoo and the subtlety of differences between Adélie and Chinstrap, consistent with the biological understanding of these species.