DEPARTMENT OF
DATA SCIENCE

DS311
Exploratory Data Analysis

Laboratory 5-8

Lab Activity: Multivariate Analysis of Penguin Data

In this lab, you will apply hierarchical clustering, K-means, PCA, and LDA to the **Palmer Penguins** dataset. This dataset contains measurements of three penguin species (*Adélie*, *Chinstrap*, *Gentoo*) in Antarctica. It's an excellent alternative to the overused iris dataset, offering a multi-faceted case study:

- **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.

Lab setup: If you have the `palmerpenguins` package, you can load the data with `data(penguins)`. If not, you can download the data from a public URL or use `install.packages("palmerpenguins")`. Ensure the data is in a data frame named `penguins` with columns `species`, `bill_length_mm`, `bill_depth_mm`, `flipper_length_mm`, `body_mass_g`, etc.

Now, work through the following exercises step by step:

1. Data Preparation: Load the penguin dataset and inspect it. How many observations and variables are there? Remove any rows with missing values. *(Hint: There are a few penguins with NA measurements which should be dropped for simplicity.)* After cleaning, report the number of penguins remaining in the data. Also, note the three species names.

2. Hierarchical Clustering: Using the four numeric features (bill length, bill depth, flipper length, body mass), perform hierarchical clustering on the penguins.

- Use **Euclidean distance** and **Ward's method** for clustering (Ward's method tends to perform well for this kind of data).

- Plot the dendrogram. Based on the dendrogram, how many clusters seem appropriate?
- Cut the dendrogram to form 3 clusters and see which species each cluster mostly corresponds to.
- Create a table of cluster assignments vs. actual species to evaluate how well unsupervised clustering recovered the true species groups.

3. K-Means Clustering: Perform K-means on the same scaled data. Since we know there are 3 species, run K-means with `k = 3`. Set a reasonable `nstart` (e.g. 20).

- Compare the K-means clusters to species labels by creating a contingency table (similar to above).
- How many penguins of each species were grouped together? Did K-means miscluster any (for example, perhaps mixing some Adelie and Chinstrap)?
- Would another choice of k perhaps make more sense from an unsupervised viewpoint? *(Bonus: try $k = 2$ or $k = 4$ and see what those clusters represent.)*

4. Principal Component Analysis: Perform PCA on the scaled feature data (bill dimensions, flipper, mass).

- How many principal components are required to explain, say, **90%** of the variance?
- Examine the loadings of the first two principal components – which variables are most influential for PC1 and PC2?
- Create a scatter plot of the penguins on the first two PC axes. Color the points by species (since we do know species in the dataset).
- Do the species form distinct clusters in PCA space? (Even though PCA didn't use species info, you can often see species separation if the inherent variation is related to species differences.)

ADVANCED QUESTION (Independent Learning)

5. Linear Discriminant Analysis: Now use LDA to build a classifier for species. Use the four numeric features as predictors.

- How well does LDA do in classifying the penguins?
- Compute the confusion matrix and overall accuracy. (You can use the training data itself for this evaluation, or do a simple train-test split if you prefer extra rigor).
- Which species pair is most frequently confused by the model? Examine the LDA coefficients for the two discriminant functions (since $G = 3$ species, there will be 2 LDs).
- What do these tell you about the differences between species? *(Hint: You might find that one LD primarily separates Gentoo penguins from the other two species, and the second LD separates Adelie vs Chinstrap.)*