# Butterlfy Species Richness

Prakhar Prakarsh
*University of Nottingham, Physics and Astronomy, Nottingham, United Kingdom*

This study employs advanced unsupervised learning techniques to analyse butterfly species distribution.Taking benefit of this rich dataset, Principal Component Analysis (PCA) and K-Means clustering were applied to unfold patterns and correlations in species richness. The analysis, grounded in robust statistical methods, reveals insightful trends in butterfly distribution, contributing significantly to ecological understanding and conservation efforts.

## I. INTRODUCTION

Studying butterfly distribution is key to ecology and conservation. Traditional techniques struggle with complex ecological data.This study uses unsupervised learning,like PCA and K-Means clustering,to simplify and analyse these complex patterns,providing fresh insights into butterfly habitats.

## II. METHODOLOGY

The analysis began by uploading the butterfly data to a DataFrame in Google Colab and exploring it preliminarily[1],which included creating a heatmap of the correlation matrix and a pair plot of the dataset.

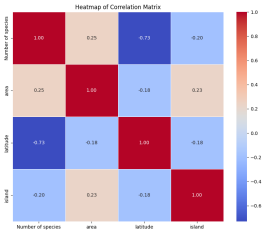


Figure 1. Correlation heatmap

The heatmap provides a detailed look at how each pair of numerical features correlates,while the pair plot gives a broader view of all relationships and distributions in the dataset.

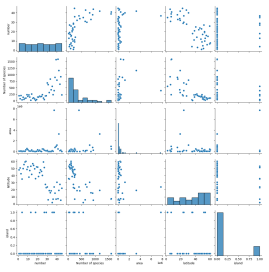

Figure 2. Pair plot/Scatterplot matrix

The data was cleaned and normalized[2],then PCA[3] reduced it to two main components.Conducting an initial analysis is crucial for identifying important patterns prior to applying PCA and clustering techniques.

$$\mathbf{PC} = \mathbf{X} \times \mathbf{W} \tag{1}$$

- **PC** represents the principal components,
- **X** is the matrix of standardized original data (after applying StandardScaler),
- **W** is the matrix of weights or loadings (eigenvectors of the covariance matrix of **X**).

The reduction made the dataset easier to interpret.K-Means clustering[4]then grouped the PCA-reduced data into three distinct clusters,

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2 \tag{2}$$

- $J$ is the objective function to be minimized,
- $k$ is the number of clusters,
- $C_i$ is the set of points in cluster $i$,
- $x$ is a data point in cluster $C_i$,
- $\mu_i$ is the centroid of the cluster $C_i$.

aimed to reveal hidden patterns and associations.The study leveraged matplotlib and seaborn libraries for visualisation[5],notably employing scatter plots to visually represent the clustering outcomes in the PCA-reduced space.

## III. CONCLUSION

The PCA application effectively condensed the dataset while retaining approximately 76.4 percent of the variance, balancing simplicity with information retention.Over 70 percent variance retention is usually enough for identifying key data patterns.Also the

heatmap shows species richness declines with latitude and slightly with islands, but increases with area. Weak correlations suggest complex ecological dynamics.

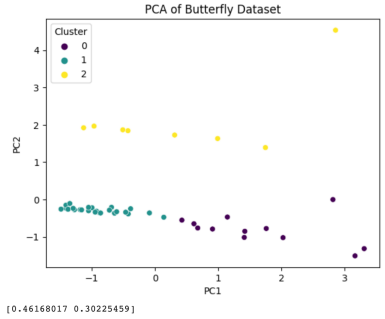


Figure 3. Variance ratio

K-Means clustering[6] revealed clear patterns in butterfly distribution. Scatter plots showed these patterns, clarifying species' spatial spread. The PCA's[3] variance ratio confirmed the dimensionality reduction's success, validating the methods used.

$$\text{Variance Ratio} = \frac{\lambda_j}{\sum_{i=1}^{n} \lambda_i} \tag{3}$$

- $\lambda_j$ is the eigenvalue corresponding to the $j^{th}$ principal component,
- $n$ is the total number of principal components.

The study confirms that unsupervised learning, through PCA and K-Means, effectively analyses ecological data and improves understanding of butterfly habitats,with broad potential for ecological research.

[1] J. W. Tukey, *Exploratory Data Analysis* (Addison-Wesley, Reading, MA, 1977).
[2] S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining* (Springer, 2015).
[3] I. T. Jolliffe, *Principal Component Analysis*, (Springer-Verlag, New York, 2002).
[4] J. B. MacQueen, *Some Methods for Classification and Analysis of Multivariate Observations*, in Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability (University of California Press, Berkeley, 1967), Vol. 1, pp. 281–297.
[5] J. VanderPlas, *Python Data Science Handbook*, (O'Reilly Media, 2016).
[6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, (Springer, 2001).