Assignment 1: Exploratory Data Analysis of a Multivariate Dataset

A. Sepúlveda-Jiménez

Data Science Dept., SoTE, CoBET, National University

TIM-8515: Multivariate Analysis

Course Instructor: Y Karahan, PhD

November 14, 2025

Contents

List of Figures

List of Tables

Introduction

This report presents a structured multivariate data analysis of the well-known Iris dataset, following the three preliminary exploratory data analysis (EDA) phases: (1) dataset exploration, (2) data cleaning (including missing value handling and outlier detection), and (3) pre-analysis multivariate visualization. Mathematical review of each section's techniques is presented. The aim is to illustrate best-practice procedures for multivariate EDA in a reusable workflow (Python pipelines). The dataset consists of 150 samples from three species of Iris flowers, each characterized by four features: sepal length, sepal width, petal length, and petal width. The analysis includes dataset exploration, data cleaning, and multivariate visualizations to explore relationships among the variables in anticipation of attempting to categorize the species of each sample.

Dataset Exploration

Let the data matrix be given by

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \tag{1}$$

where in our case $n = 150$ observations and $p = 4$ continuous variables (sepal length, sepal width, petal length, petal width). Additionally, we have a categorical response vector $\mathbf{y} \in \{1, 2, 3\}^n$ corresponding to three possible species. Use of simultaneous analysis of all p variables falls under the umbrella of multivariate data analysis (MDA) (Dempster, 1971; Everitt, 1975; Hazra & Gogtay, 2017; McQuitty, 2018).

We compute the sample mean vector and sample covariance matrix:

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \quad , \quad \mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\top}. \tag{2}$$

Here $\mathbf{x}_i$ is the i-th row of $\mathbf{X}$. The correlation matrix is

$$\mathbf{R} = \mathrm{diag}(\mathbf{S})^{-1/2}\,\mathbf{S}\,\mathrm{diag}(\mathbf{S})^{-1/2}. \tag{3}$$

We inspect $\mathbf{R}$ to assess pairwise relationships among the continuous variables. For instance, the correlation between petal length and petal width is high (≈ 0.96) in this dataset. A covariance scatter plot between two features would show a ellipse-like pattern for data that more closely follow a bivariate normal distribution. The ellipse would tilt upwards from bottom-left to top-right when the covariance is positive and be tilted downwards from top-left to bottom-right if the covariance is negative. Furthermore, the shape or strength of correlation would generate a thiner ellipse for strong correlations and fatter, nearly circular one, for weak/zero correlations. See Figure!1 for a schematic example of this.

Variable types and missing values

All four feature variables are continuous; the target is categorical with three levels. There are no missing values in the dataset (i.e., $\sum_i \mathbf{1}[x_{ij} \text{ is missing}] = 0\ \forall j$). We therefore proceed without imputation but still show the generic imputation logic in Section 3.

Correlation analysis

The correlation matrix (heatmap) reveals strong positive associations among petal length and petal width, moderate correlation between sepal length and petal length, and weaker or negative correlations involving sepal width. See Figure 6. The structure of $\mathbf{R}$ justifies the use of methods that exploit multivariate dependencies (see e.g. (Adachi & van de Velden, 2025)).

The dataset is composed of 150 observations, with 4 continuous features and a categorical target variable, which represents the species of the Iris flower. The features are as follows:

- Sepal Length (cm)
- Sepal Width (cm)
- Petal Length (cm)
- Petal Width (cm)

The target variable is the species, which takes one of the following category values:

- Setosa
- Versicolor

- Virginica

Data Summary

The dataset has no missing values. The following summary statistics for each feature were observed:

Correlation Matrix

The correlation matrix (heatmap) for the numerical features of the dataset is shown in Figure fig:heatmap. Strong positive correlations were observed between petal length and petal width, as well as between sepal length and petal length.

Data Cleaning

Given the absence of missing values, the focus shifts to detection of outliers. We standardise each variable:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \quad j = 1, \dots, p, \ i = 1, \dots, n, \tag{4}$$

where $s_j = \sqrt{S_{jj}}$. Observations for which $|z_{ij}| > 3$ (a conventional threshold) on any dimension are flagged as outliers. In this dataset, the count of flagged observations is minimal (often zero), so no removal was executed. The general decision rule is:

$$\mathbf{1}\Big(\max_j |z_{ij}| > 3\Big) = 1 \ \Rightarrow \ \text{“suspected outlier”}. \tag{5}$$

We show a generic schematic for how a box-and-whisker diagram would visually present with outliers and the general spread of the data. See Figure 2 . Imputation for missing-value cases would proceed via

$$x_{ij}^{\text{imp}} = \bar{x}_j \tag{6}$$

or more advanced techniques (e.g., EM-based) if missingness is non-ignorable (see (Everitt, 1975; McQuitty, 2018)).

No missing values were found in the dataset. However, outlier detection was performed using the Z-score method. Observations with Z-scores greater than 3 were considered outliers. No extreme outliers were detected in the dataset.

Generalized Principal Components

To reduce the dimensionality of the dataset and observe the main axes of variation, we perform a PCA. Let $\mathbf{X}_{\text{std}} \in \mathbb{R}^{n \times p}$ denote the standardized features *column-centered* data matrix with rows $\mathbf{x}_i^\top$ $(i = 1, \ldots, n)$. Define the sample covariance estimator

$$\widehat{\boldsymbol{\Sigma}} \;=\; \frac{1}{n}\, \mathbf{X}^\top \mathbf{X} \in \mathbb{R}^{p \times p}. \tag{7}$$

(Some software uses $1/(n-1)$; this rescales eigenvalues but not eigenvectors.)

Variance–maximization view. The first principal direction $\mathbf{v}_1 \in \mathbb{R}^p$ maximizes the variance of the projected data subject to unit length:

$$\mathbf{v}_1 \;=\; \arg \max_{\|\mathbf{v}\|_2 = 1} \operatorname{Var}\Big(\mathbf{X}\mathbf{v}\Big) \;=\; \arg \max_{\|\mathbf{v}\|_2 = 1} \mathbf{v}^\top \widehat{\boldsymbol{\Sigma}}\, \mathbf{v}. \tag{8}$$

See Figure 3 for a schematic of a PCA component plot. Introducing a Lagrange multiplier for the constraint yields the eigenvalue problem $\widehat{\boldsymbol{\Sigma}}\, \mathbf{v} = \lambda\, \mathbf{v}$; hence $\mathbf{v}_1$ is the leading eigenvector and λ_1 its associated eigenvalue. Subsequent directions $\mathbf{v}_m$ $(m = 2, \ldots, p)$ solve the same program with the additional orthogonality constraints $\mathbf{v} \perp \{\mathbf{v}_1, \ldots, \mathbf{v}_{m-1}\}$. Let $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_p]$ and $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ with $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$. Then

$$\widehat{\boldsymbol{\Sigma}} \;=\; \mathbf{V}\, \boldsymbol{\Lambda}\, \mathbf{V}^\top, \qquad \mathbf{v}_m = m\text{th eigenvector}, \quad \lambda_m = \text{variance of component } m. \tag{9}$$

The principal component (PC) scores are

$$\mathbf{z}_m \;=\; \mathbf{X}\, \mathbf{v}_m \in \mathbb{R}^n, \qquad \operatorname{Var}(\mathbf{z}_m) = \lambda_m, \quad \operatorname{Cov}(\mathbf{z}_m, \mathbf{z}_\ell) = 0 \;\; (m \neq \ell). \tag{10}$$

The fraction of variance explained (PVE) by the first M components is

$$\operatorname{PVE}(M) \;=\; \frac{\sum_{m=1}^{M} \lambda_m}{\sum_{j=1}^{p} \lambda_j}. \tag{11}$$

SVD connection. Let the thin singular value decomposition (SVD) be

$$\mathbf{X} \;=\; \mathbf{U}\,\mathbf{D}\,\mathbf{V}^{\top}, \qquad \mathbf{U} \in \mathbb{R}^{n\times r},\ \mathbf{V} \in \mathbb{R}^{p\times r},\ \mathbf{D} = \operatorname{diag}(d_1, \dots, d_r),\ r = \operatorname{rank}(\mathbf{X}). \tag{12}$$

Then

$$\widehat{\mathbf{\Sigma}} \;=\; \frac{1}{n}\mathbf{X}^{\top}\mathbf{X} \;=\; \mathbf{V}\,\frac{\mathbf{D}^2}{n}\,\mathbf{V}^{\top}, \tag{13}$$

so principal directions are the columns of $\mathbf{V}$ and

$$\lambda_m \;=\; \frac{d_m^2}{n}, \qquad \mathbf{z}_m \;=\; \mathbf{X}\mathbf{v}_m \;=\; d_m\,\mathbf{u}_m. \tag{14}$$

Collecting the first M components gives $\mathbf{Z}_M = \mathbf{X}\mathbf{V}_M = \mathbf{U}_M\mathbf{D}_M$ with $\mathbf{V}_M = [\mathbf{v}_1, \dots, \mathbf{v}_M]$.

Best rank-M reconstruction. Truncating the SVD provides the least-squares optimal rank-M approximation

$$\mathbf{X}_M \;=\; \arg\min_{\operatorname{rank}(\mathbf{A})\le M} \;\|\mathbf{X} - \mathbf{A}\|_F^2 \;=\; \mathbf{U}_M\,\mathbf{D}_M\,\mathbf{V}_M^{\top}, \tag{15}$$

equivalently $\widehat{\mathbf{X}} = \mathbf{X}\mathbf{V}_M\mathbf{V}_M^{\top}$ with residual error $\sum_{m>M} d_m^2$.

Scaling and correlation PCA.. If features differ in physical units or scale, standardize columns of $\mathbf{X}$ (or, equivalently, perform PCA on the correlation matrix $\mathbf{R} = \operatorname{diag}(\widehat{\mathbf{\Sigma}})^{-1/2}\,\widehat{\mathbf{\Sigma}}\,\operatorname{diag}(\widehat{\mathbf{\Sigma}})^{-1/2}$) to prevent domination by high-variance coordinates.

The presentation above follows the treatment in (Hastie et al., 2009). Figure 8 illustrates the actual Iris dataset's first two principal components.

Multivariate Visualization

We employ four primary visualisation tools: (i) pairwise scatter-plots (pair plots) color-coded by species, (ii) heatmap of the correlation matrix, (iii) box-whisker plots to show the data spread of each feature and category variable, and (iv) PCA plot to show the principal component breakdown of the dataset, the most relevant dependencies, and the scatter of category classes. Visualization is essential in multivariate analysis to detect clusters, separation by groups, and multicollinearity, as emphasized in (Hazra & Gogtay, 2017; McQuitty, 2018). Figure 7 shows the pair-plot; Figure 6 shows the correlation heatmap.

Additional Multivariate Visualizations: Parallel Coordinates, 3D Scatter, and t-SNE

Beyond pairwise scatterplots and correlation heatmaps, several additional visualization techniques are useful for understanding the joint structure of the Iris data in higher dimensions. Here we comment on three such displays: the parallel coordinates plot, the 3D scatter plot, and a non-linear embedding via t-distributed stochastic neighbor embedding (t-SNE). Each of these views highlights complementary aspects of the same underlying data and is widely used in exploratory multivariate analysis (Cleveland, 1993; van der Maaten & Hinton, 2008; Wegman, 1990).

Parallel Coordinates Plot

In a parallel coordinates plot, each feature is represented as a vertical axis; every observation is drawn as a polyline that intersects each axis at the value of the corresponding feature. This transforms a p-dimensional point $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ into a polygonal chain in $\mathbb{R}^2$, while still allowing one to see multivariate patterns across all p dimensions at once (Wegman, 1990).

Figure 9 shows the parallel coordinates plot for the four numeric Iris features, with lines colored by species. Several patterns are immediately visible:

- *Setosa* lines cluster tightly at low petal length and petal width, clearly separated from the other species on those axes.
- *Versicolor* and *virginica* overlap more, but their typical trajectories differ: virginica tends to have larger petal length and petal width, and slightly larger sepal length.
- Crossings of lines between axes (e.g., between sepal width and petal length) visually encode negative or weak correlations across species.

These patterns are consistent with the correlation matrix and PCA results, but the parallel coordinate view emphasizes individual observations and makes it easy to spot unusual profiles (potential outliers or mixed cases).

3D Scatter Plot

While ordinary scatterplots display two dimensions at a time, a 3D scatter plot visualizes three features simultaneously, using spatial position in $\mathbb{R}^3$ to encode the triplet $(x_{ij_1}, x_{ij_2}, x_{ij_3})$. Figure 10 shows a 3D scatter plot based on standardized sepal length, sepal width, and petal length, with points colored by species.

Several aspects are worth noting:

- The setosa cluster is again compact and well-separated in this 3D space, reinforcing that even simple linear combinations of these three features can distinguish it from the other two species.
- Versicolor and virginica are partially overlapped but occupy different regions along the petal-length direction, hinting at a nearly linear separation between them that PCA and linear classification can exploit.
- Rotating the 3D view (interactively, in a notebook or visualization tool) often reveals directions in which clusters appear most distinct, which is conceptually aligned with the PCA idea of finding variance-maximizing projections (Cleveland, 1993; Hastie et al., 2009).

t-SNE Embedding

Both parallel coordinates and 3D scatter plots use (approximately) linear geometry: they show the data in either the original feature space or in simple linear projections. t-distributed stochastic neighbor embedding (t-SNE) instead constructs a non-linear, low-dimensional embedding that attempts to preserve local neighborhood structure of the high-dimensional data (van der Maaten & Hinton, 2008).

At a high level, t-SNE defines pairwise similarities between points in the original space using (symmetrized) Gaussian kernels, then learns two-dimensional coordinates $\{\mathbf{y}_i\}_{i=1}^{n}$ such that similar points in the original space remain close in the embedding while dissimilar points are modeled far apart. This is achieved by minimizing a Kullback–Leibler divergence between the high-dimensional and low-dimensional similarity distributions.

Figure 11 shows a 2D t-SNE embedding of the standardized Iris features. In this view:

- Each species forms a tight cluster, with very little overlap.
- The relative distances between clusters are *not* directly interpretable as metric distances (t-SNE preserves local neighborhoods better than global geometry), but the clear separation of groups confirms that the class structure is strong and low-dimensional.
- Within-cluster shape can sometimes hint at substructure (e.g., curvature or subclusters), which may be useful in more complex datasets.

In combination, the parallel coordinates plot, 3D scatter plot, and t-SNE embedding provide a richer understanding of the Iris data than any single visualization alone: parallel coordinates emphasize high-dimensional profiles, 3D scatter retains metric structure in three selected directions, and t-SNE uncovers non-linear clustering in a purely two-dimensional display.

Conclusion

This report carried out a multivariate exploratory analysis of the Iris dataset with an emphasis on understanding both marginal behaviour of the features and their joint relationships across species. The workflow combined classical numerical summaries with a suite of complementary visual tools, including box-and-whisker plots, correlation heatmaps, pairwise scatterplots, principal component analysis (PCA) and its biplot, parallel–coordinates plots, three–dimensional scatter plots, and a non–linear t-SNE embedding.

From a data-quality perspective, there were no missing values and essentially no extreme outliers. The univariate summaries and boxplots showed that the three species have markedly different distributions for petal length and petal width, whereas sepal dimensions are less discriminative. The correlation matrix and its heatmap revealed strong positive association between petal length and petal width and more modest relationships involving sepal width, suggesting that most of the intrinsic variation and class separation is driven by the petal measurements.

PCA formalized this picture. On the standardized features, the first two principal components captured roughly 96% of the total variance, and their loadings were dominated by the petal variables. The PCA biplot showed setosa as a well-separated cluster and indicated that versicolor and virginica are mainly separated along a direction associated with increasing petal length and width. This confirmed that the effective intrinsic dimension of the dataset, from a variance and classification point of view, is low (two to three dimensions).

The additional multivariate visualizations refined this interpretation. The parallel–coordinates plot made species-specific “profiles” across all four features visible in a single display: setosa forms a tight bundle with uniformly small petal values, whereas versicolor and virginica show overlapping but systematically shifted trajectories, particularly on the petal axes. The 3D scatter plot, using selected standardized features, reinforced that setosa occupies a compact, isolated region, while the other two species differ along an oblique direction combining sepal and petal information. Finally, the t-SNE embedding produced three distinct clusters in two dimensions, confirming that the class structure is genuinely strong and that a low-dimensional representation can faithfully capture local neighborhoods, even if the t-SNE axes themselves are not directly interpretable.

Taken together, these results paint a consistent picture:

- The Iris dataset is clean, low-dimensional, and exhibits clear multivariate structure dominated by petal length and petal width.

- Setosa is easily separable from the other two species in almost any reasonable projection, while the versicolor–virginica separation is moderate and depends on combining several features.

- Linear methods (such as PCA-based dimension reduction and simple linear classifiers) are likely sufficient to achieve high accuracy, but more flexible decision boundaries can be justified if the goal is to finely separate versicolor from virginica.

In practice, this analysis illustrates how a sequence of complementary multivariate plots—from boxplots and scatterplot matrices to PCA biplots, parallel coordinates, 3D views, and t-SNE—can be used to build a coherent understanding of the structure of a real dataset before any formal modelling is attempted. The same workflow can be transferred to higher-dimensional and less well-behaved data, where the stakes of getting the geometry wrong are much higher than in this pedagogical example.

References

Adachi, K., & van de Velden, M. (2025). Advances in multivariate data analysis. *Behaviormetrika*, *52*, 413–415.

Borsboom, D., Deserno, M. K., & Waldorp, L. J. (2021). Network analysis of multivariate data in psychological science. *Nature Reviews Methods Primers*, *1*(1), 58.

Cleveland, W. S. (1993). *Visualizing data*. Hobart Press.

Dempster, A. P. (1971). An overview of multivariate analysis [emphasis on multivariate methods in practice]. *Journal of Multivariate Analysis*, *1*(1), 1–20.

Everitt, B. S. (1975). Multivariate analysis: The need for data, and other problems. *British Journal of Psychiatry*, *126*(3), 237–240.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Hazra, A., & Gogtay, N. (2017). Biostatistics series module 10: Brief overview of multivariate methods. *Indian Journal of Dermatology*, *62*(4), 358–366. https://doi.org/10.4103/ijd.IJD_296_17

McQuitty, S. (2018). The purposes of multivariate data analysis methods. *Measurement and Evaluation in Counseling and Development*, *51*(1), 46–53.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*, 2579–2605.

Wegman, E. J. (1990). Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, *85*(411), 664–675. https://doi.org/10.1080/01621459.1990.10474926

Appendix: Tables and Figures

Table 1

Summary Statistics of the Iris Dataset

Feature	**Mean**	**Std**	**Min**	**Max**	**Range**
Sepal Length (cm)	5.84	0.83	4.30	7.90	3.60
Sepal Width (cm)	3.05	0.43	2.00	4.40	2.40
Petal Length (cm)	3.76	1.76	1.00	6.90	5.90
Petal Width (cm)	1.20	0.76	0.10	2.50	2.40

Table 2

Outlier Detection (Z-score > 3)

Feature	**Outliers (count)**
Sepal Length (cm)	0
Sepal Width (cm)	0
Petal Length (cm)	0
Petal Width (cm)	0

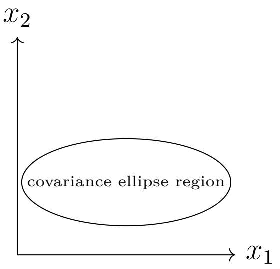

Figure 1

Schematic covariance-ellipse in two dimensions (example diagram).

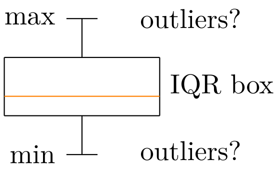

Figure 2

Schematic box-and-whisker representation for outlier identification (example diagram).

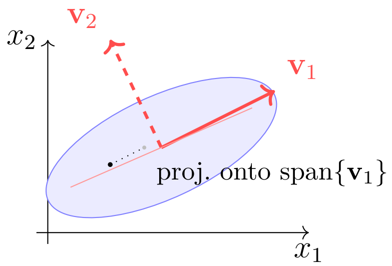

Figure 3

Geometric picture: centered data (ellipse) projected onto the first principal direction $\mathbf{v}_1$*, which maximizes variance of* $\mathbf{X}\mathbf{v}$*. The dashed arrow shows the orthogonal second direction* $\mathbf{v}_2$*.*

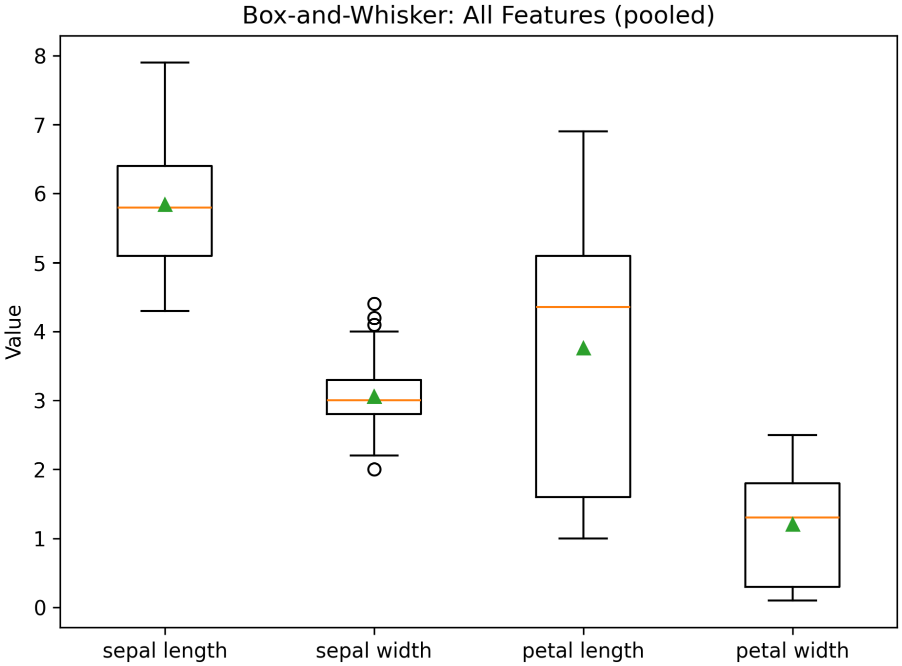

Figure 4

Box-and-whisker plots for all features (pooled across species).

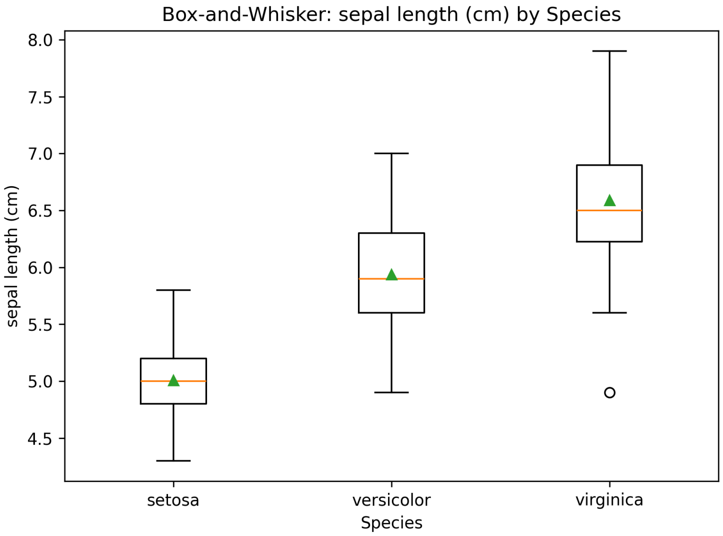

Figure 5

Box-and-whisker plot of sepal length by species.

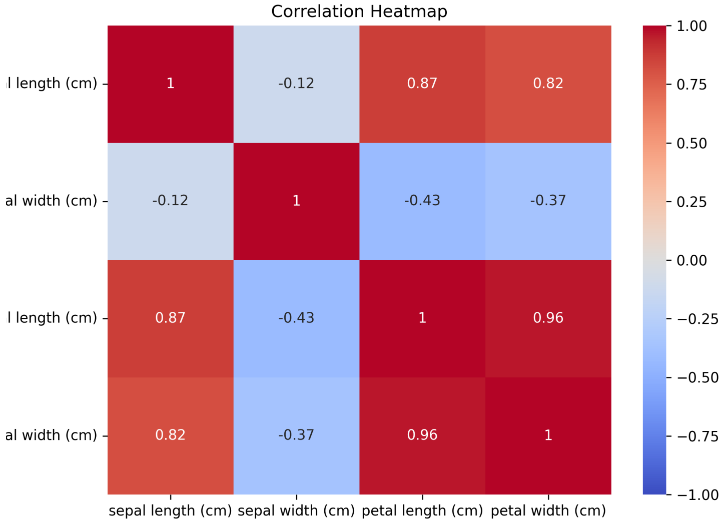

Figure 6

Heatmap of the correlation matrix $\mathbf{R}$.

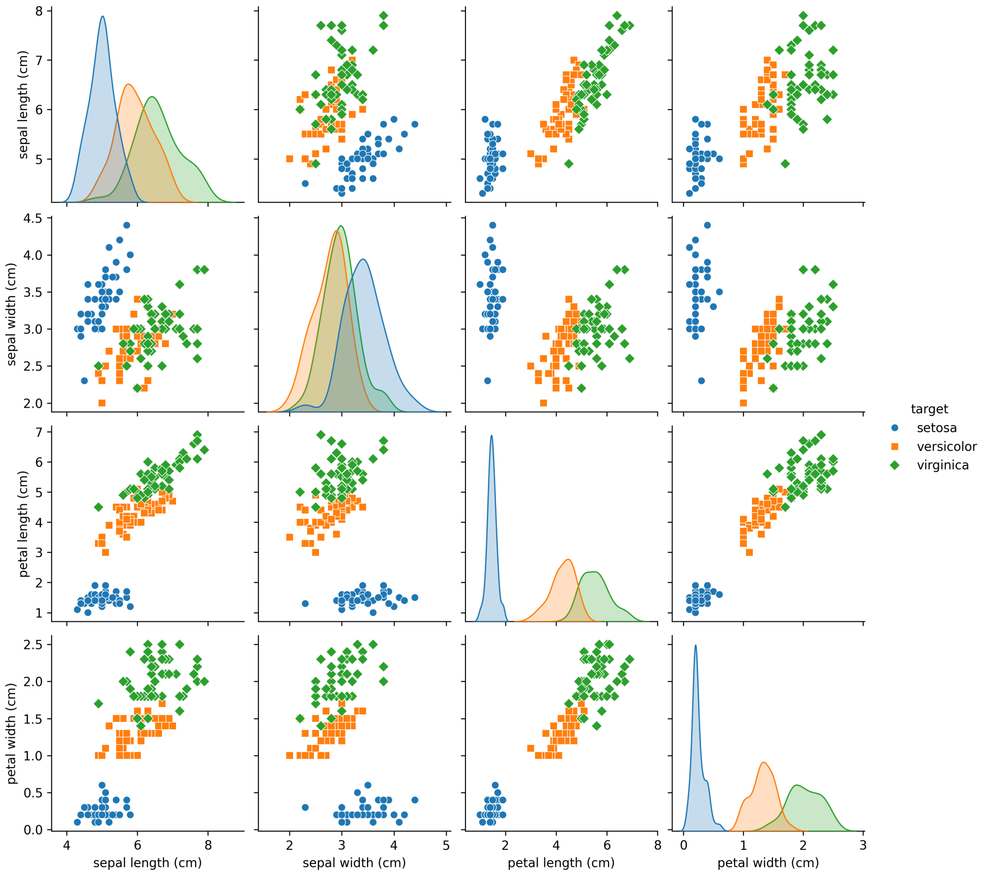

Figure 7

Pair-plot of the four continuous features of the Iris dataset, colour-coded by species.

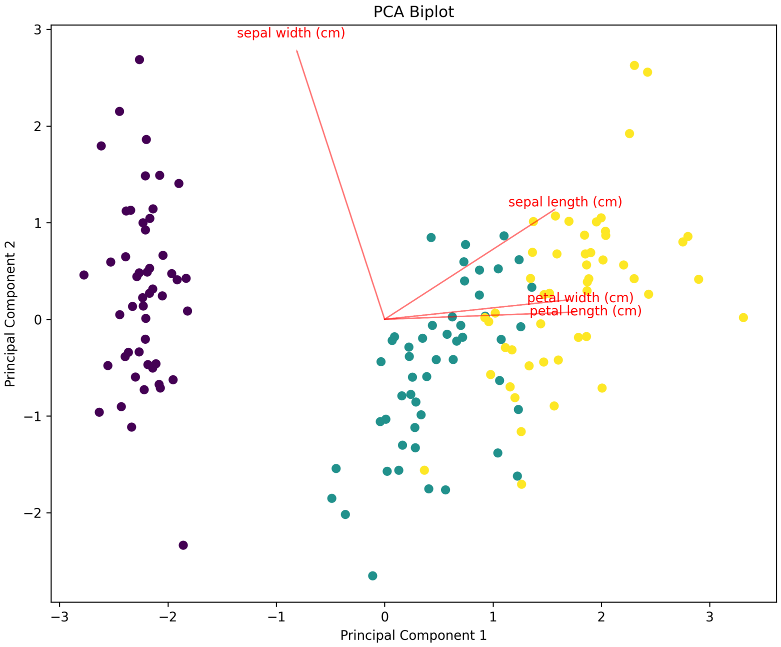

Figure 8

PCA biplot showing the first two principal components.

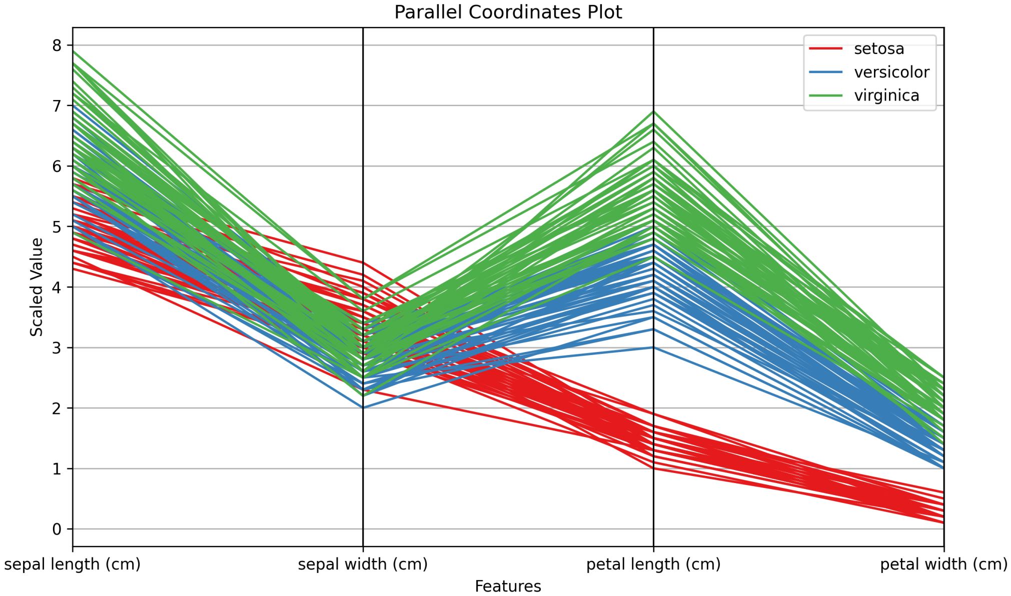

Figure 9

Parallel coordinates plot for the four Iris features, with each line representing one flower and colored by species. Distinct bundles of lines correspond to species-specific profiles across all features.

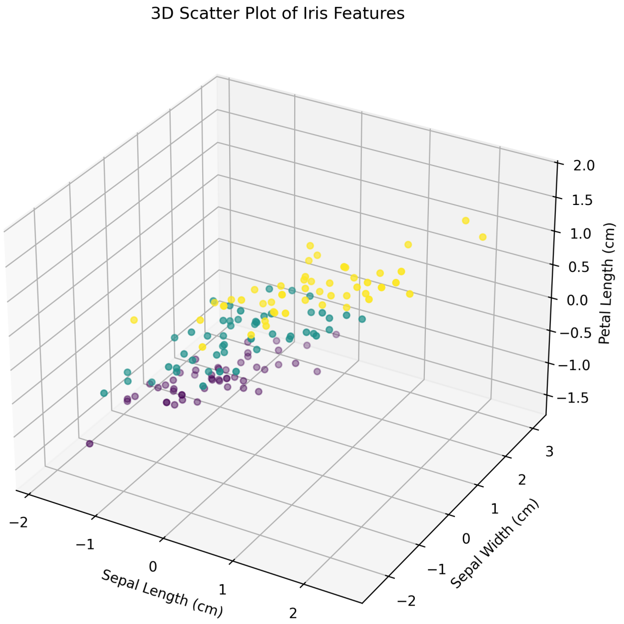

Figure 10

3D scatter plot of standardized sepal length, sepal width, and petal length for the Iris data. Points are colored by species. The setosa cluster is well separated; versicolor and virginica show partial overlap with structure along the petal-length direction.

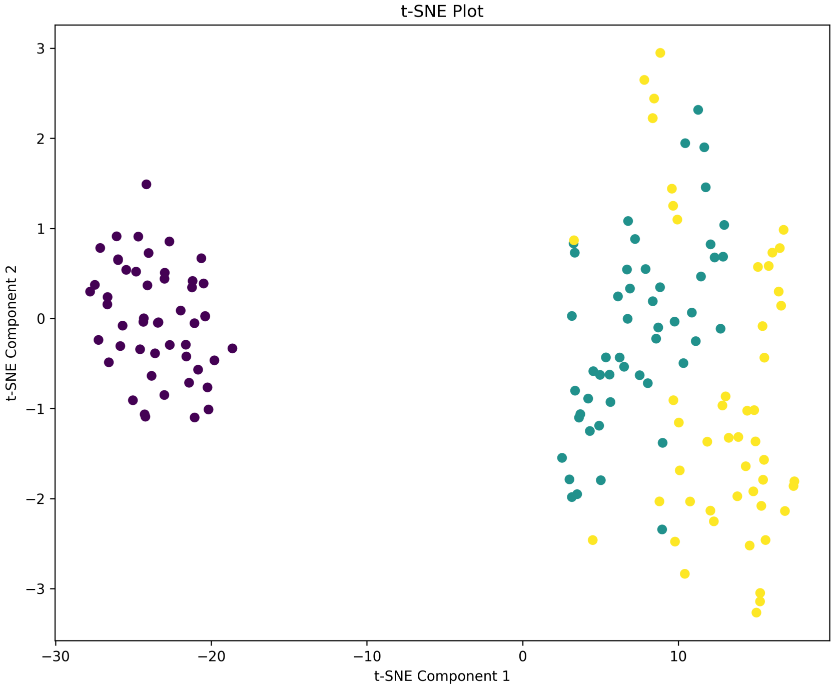

Figure 11

Two-dimensional t-SNE embedding of the standardized Iris features. Each point corresponds to one flower, colored by species. t-SNE reveals three well-separated clusters, reflecting strong class structure in the original feature space.