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 11, 2025

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

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

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:

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

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

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

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.

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:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \quad j = 1, \ldots, p, \; i = 1, \ldots, n, \tag{3}$$

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:

$$\mathbf{1}\Big(\max_j |z_{ij}| > 3\Big) = 1 \;\Rightarrow\; \text{“suspected outlier”}. \tag{4}$$

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

$$x_{ij}^{\text{imp}} = \bar{x}_j \tag{5}$$

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

$$\widehat{\mathbf{\Sigma}} \;=\; \frac{1}{n}\,\mathbf{X}^\top \mathbf{X} \in \mathbb{R}^{p \times p}. \tag{6}$$

(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:

$$\mathbf{v}_1 \;=\; \arg\max_{\|\mathbf{v}\|_2 = 1} \operatorname{Var}\Big(\mathbf{X}\mathbf{v}\Big) \;=\; \arg\max_{\|\mathbf{v}\|_2 = 1} \; \mathbf{v}^\top \widehat{\mathbf{\Sigma}}\, \mathbf{v}. \tag{7}$$

Introducing a Lagrange multiplier for the constraint yields the eigenvalue problem $\widehat{\mathbf{\Sigma}}\,\mathbf{v} = \lambda\,\mathbf{v}$; 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 $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ with $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$. Then

$$\widehat{\mathbf{\Sigma}} \;=\; \mathbf{V}\,\mathbf{\Lambda}\,\mathbf{V}^\top, \qquad \mathbf{v}_m = m\text{th eigenvector}, \quad \lambda_m = \text{variance of component } m. \tag{8}$$

The principal component (PC) scores are

$$\mathbf{z}_m \;=\; \mathbf{X}\,\mathbf{v}_m \in \mathbb{R}^n, \qquad \operatorname{Var}(\mathbf{z}_m) = \lambda_m, \quad \operatorname{Cov}(\mathbf{z}_m, \mathbf{z}_\ell) = 0 \;\; (m \neq \ell). \tag{9}$$

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

$$\text{PVE}(M) \;=\; \frac{\sum_{m=1}^{M} \lambda_m}{\sum_{j=1}^{p} \lambda_j}. \tag{10}$$

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

$$\mathbf{X} \;=\; \mathbf{U}\,\mathbf{D}\,\mathbf{V}^\top, \qquad \mathbf{U} \in \mathbb{R}^{n \times r},\; \mathbf{V} \in \mathbb{R}^{p \times r},\; \mathbf{D} = \operatorname{diag}(d_1, \ldots, d_r),\; r = \operatorname{rank}(\mathbf{X}). \tag{11}$$

Then

$$\widehat{\mathbf{\Sigma}} \;=\; \frac{1}{n}\,\mathbf{X}^\top \mathbf{X} \;=\; \mathbf{V}\,\frac{\mathbf{D}^2}{n}\,\mathbf{V}^\top, \tag{12}$$

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

$$\lambda_m \;=\; \frac{d_m^2}{n}, \qquad \mathbf{z}_m \;=\; \mathbf{X}\mathbf{v}_m \;=\; d_m\,\mathbf{u}_m. \tag{13}$$

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

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

$$\mathbf{X}_M \;=\; \arg\min_{\operatorname{rank}(\mathbf{A})\le M} \|\mathbf{X}-\mathbf{A}\|_F^2 \;=\; \mathbf{U}_M\,\mathbf{D}_M\,\mathbf{V}_M^\top, \tag{14}$$

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 principlal 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.

Conclusion

This analysis demonstrates a straightforward, yet rigorous, workflow for multivariate exploratory data analysis on a real dataset. We documented the structure of the data, handled potential issues of missing values and outliers, and produced

visualizations that highlight inter-variable relationships and potential group separations. We appled dimension-reduction method via an PCA. Classification methods (LDA, logistic regression), or network-based approaches to multivariate dependencies may then follow for a more precise relationship and classification analysis (Borsboom et al., 2021). Overall, the process adheres to best-practices described in the multivariate statistics literature (Adachi & van de Velden, 2025; Everitt, 1975; McQuitty, 2018). This basic analysis of the Iris dataset provided insights into the relationships among features and species. The dataset did not contain missing values, and no significant outliers were detected. The pairplot and heatmap visualizations helped reveal interesting patterns and correlations between the features, which may be useful for further modeling.

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.

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.

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

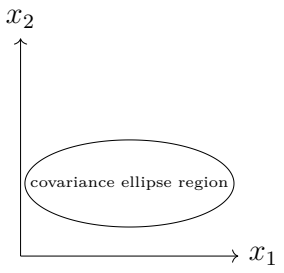

Figure 1

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

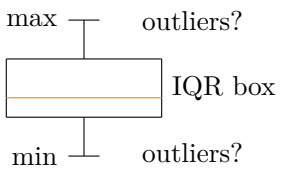

Figure 2

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

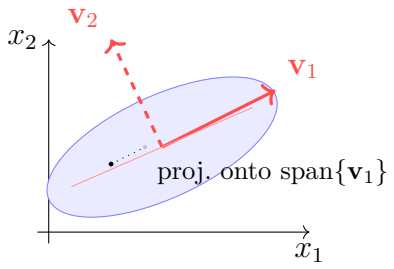

Figure 3

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

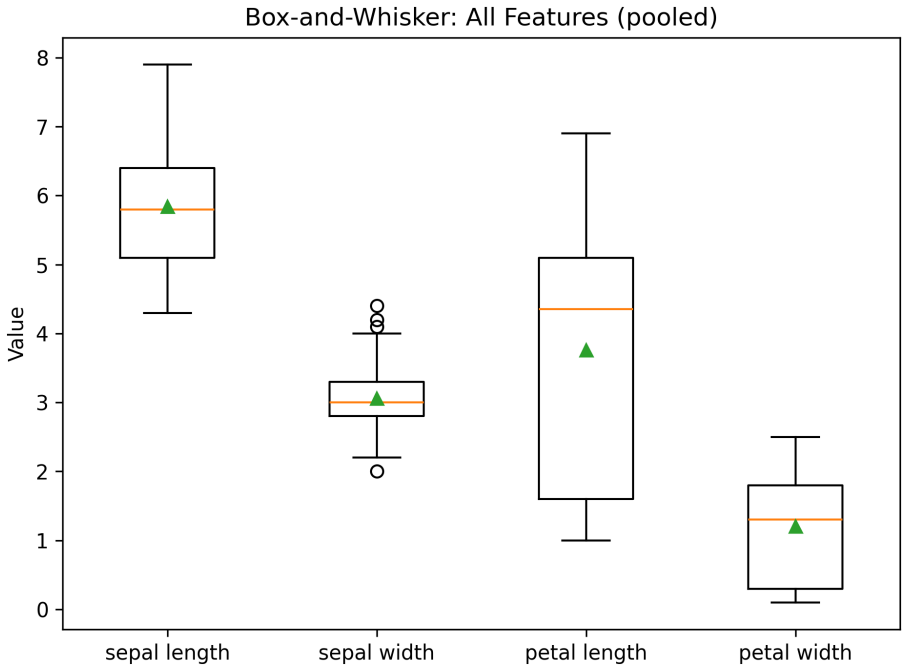

Figure 4

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

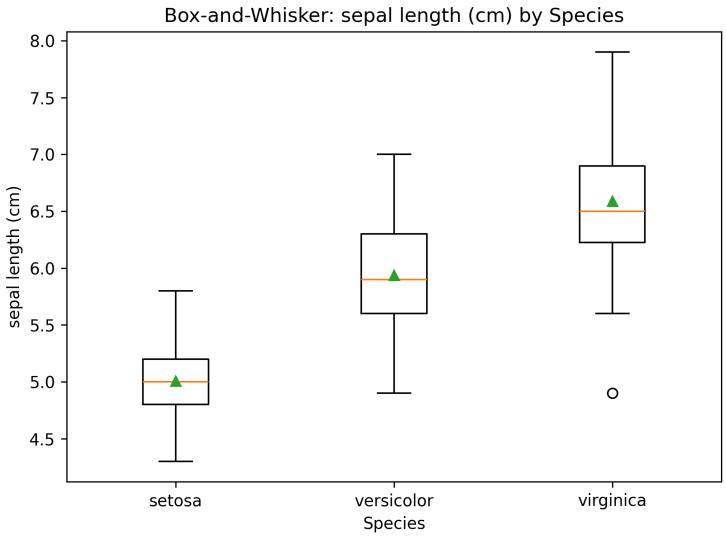

Figure 5

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

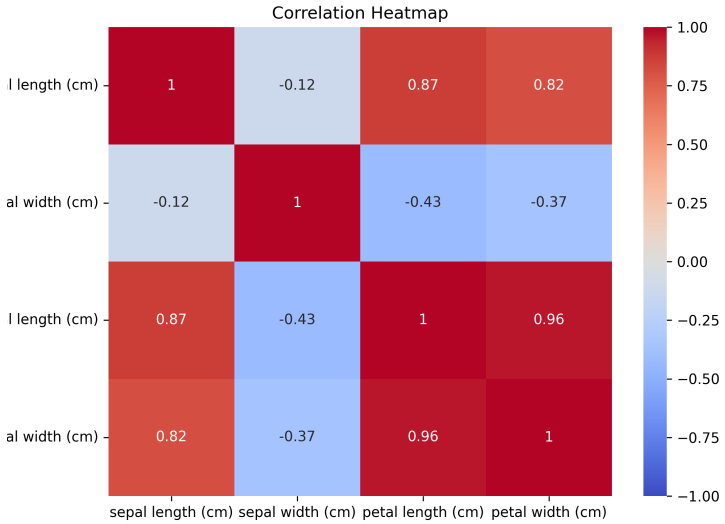

Figure 6

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

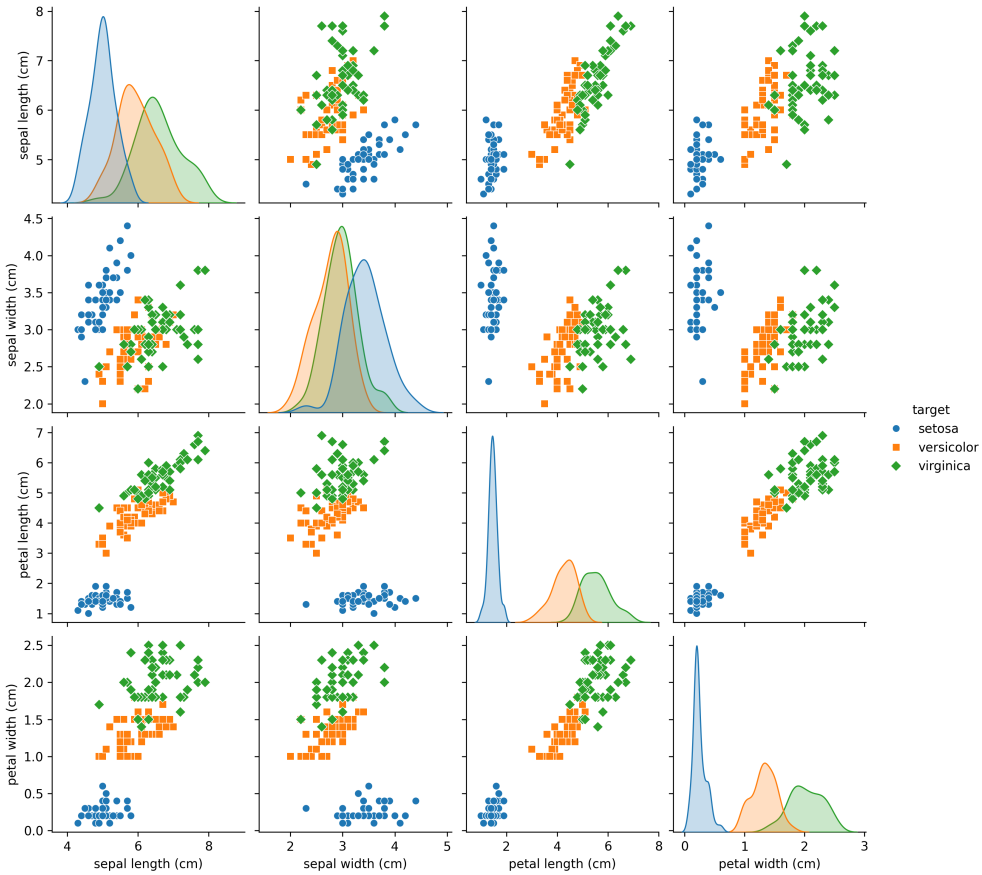

Figure 7

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

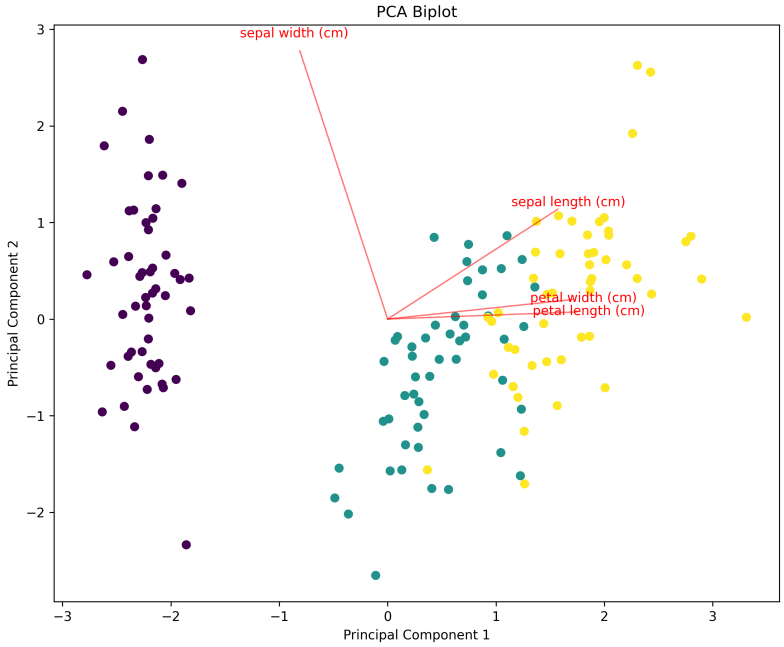

Figure 8

PCA plot showing the first two principal components.