Course: 2025 Fall 1 CSC 7900 for Andrew Webb
Student: Roohana Karim
Date: 9/28/2025
Dataset: *iris* (Fisher, 1936)

1. **Dataset Description**

The *iris* dataset contains 150 flower measurements across 3 species (*setosa, versicolor, virginica*). Variables include Sepal.Length, Sepal.Width, Petal.Length, Petal.Width (cm) and a categorical Species label (balanced: 50 each). There are no missing values.

Goal. Explore multivariate structure, reduce dimensionality (PCA), discover groups (clustering), and model Sepal.Length via regression.

'data.frame':150 obs. of 5 variables:

$ Sepal.Length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...

$ Sepal.Width : num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...

$ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...

$ Petal.Width : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...

$ Species : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...

Summary:

Sepal.Length	*Sepal.Width*	*Petal.Length*	*Petal.Width*
Min. :4.300	*Min. :2.000*	*Min. :1.000*	*Min. :0.100*
1st Qu.:5.100	*1st Qu.:2.800*	*1st Qu.:1.600*	*1st Qu.:0.300*
Median :5.800	*Median :3.000*	*Median :4.350*	*Median :1.300*
Mean :5.843	*Mean :3.057*	*Mean :3.758*	*Mean :1.199*

3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800

Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500

Species

setosa :50

versicolor:50

virginica :50

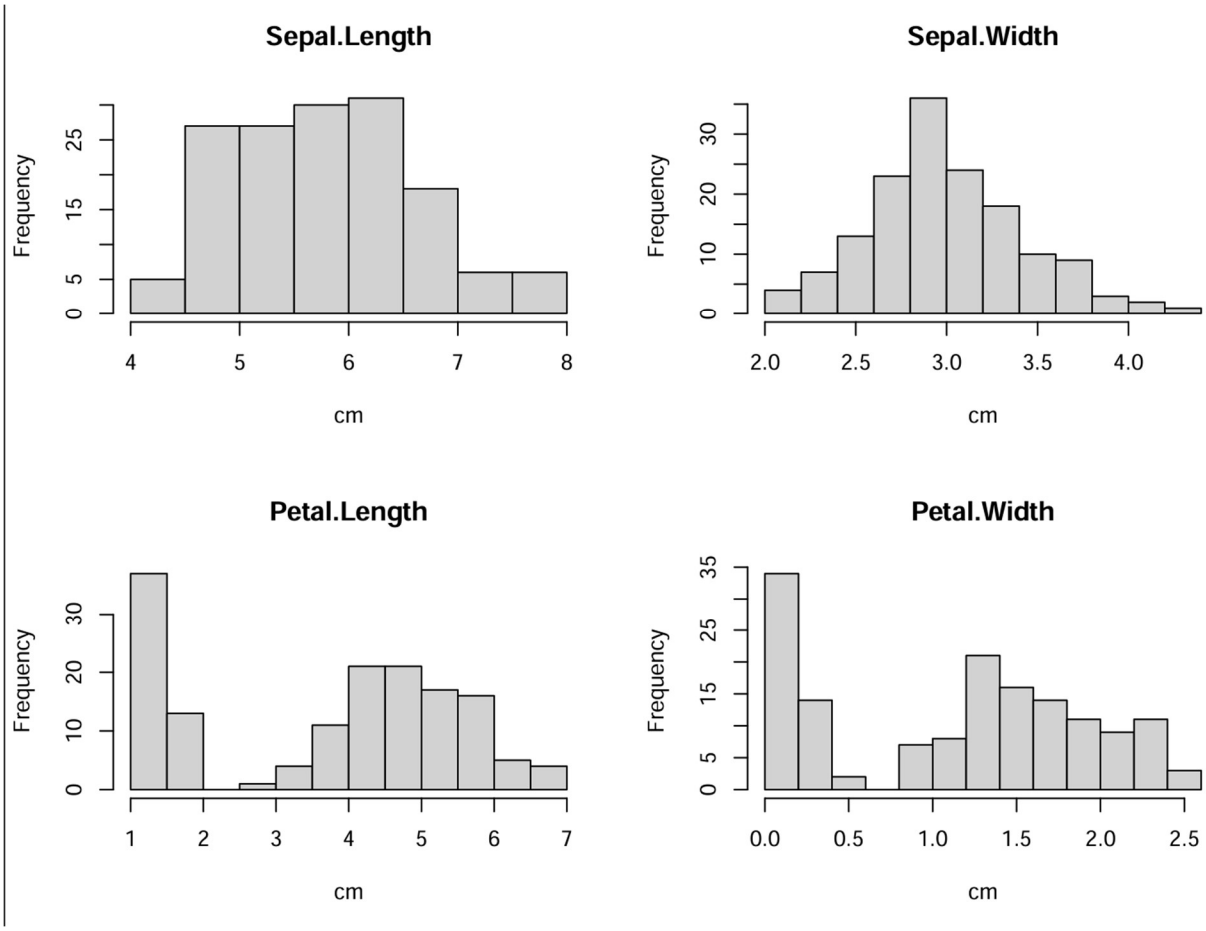
Missing values per column:

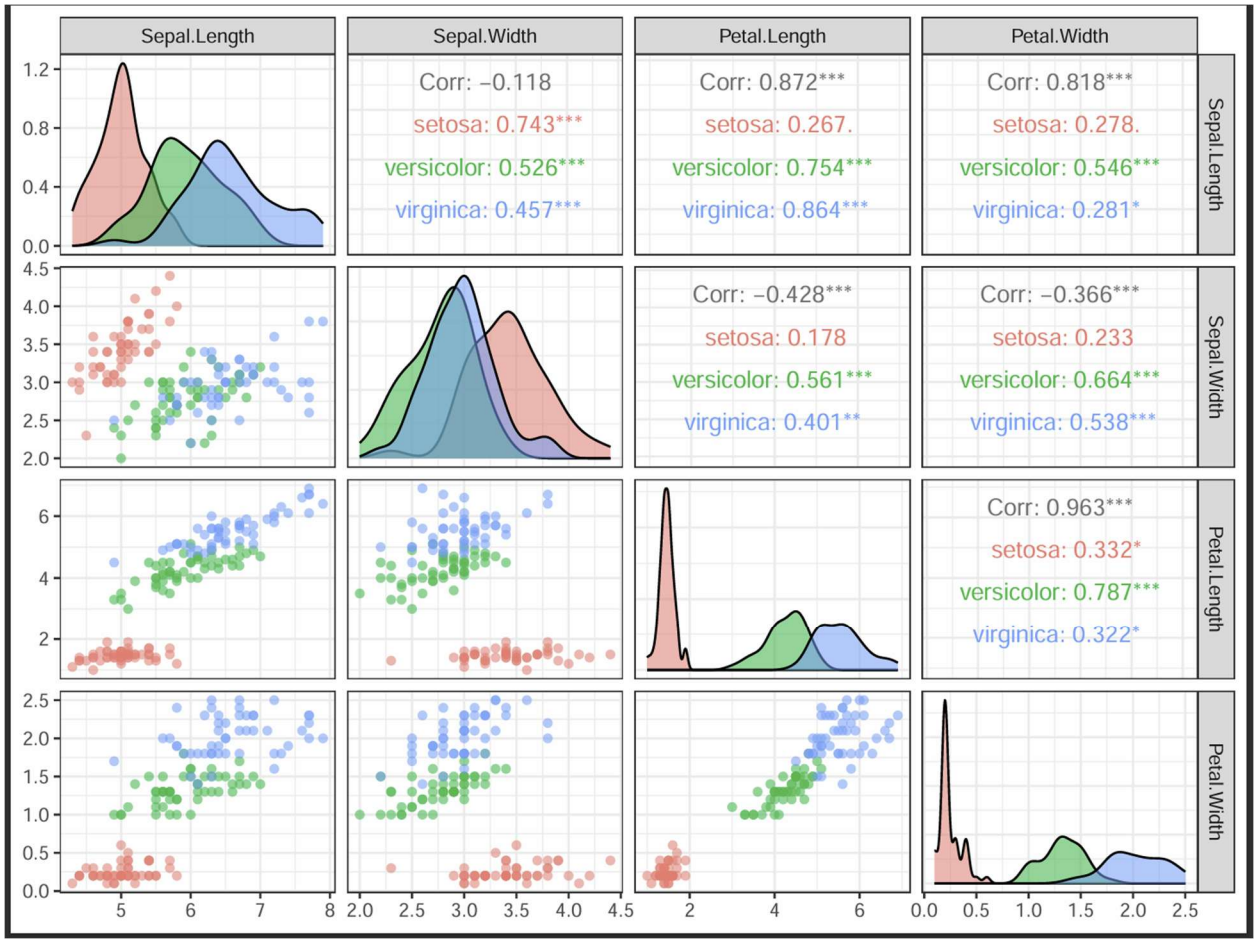
Sepal.Length	*Sepal.Width*	*Petal.Length*	*Petal.Width*	*Species*
0	*0*	*0*	*0*	*0*

2. Data Preprocessing & Exploratory Analysis (EDA)

All four numeric variables were standardized (mean 0, sd 1) before PCA and clustering.
Distributions show different spreads for sepal vs petal measures. Correlations reveal strong associations among petal measures and between petal length and sepal length:

- Cor(Petal.Length, Petal.Width) ≈ 0.963
- Cor(Sepal.Length, Petal.Length) ≈ 0.872
- Sepal.Width is negatively correlated with petal variables.


Sepal.Length
Frequency
0
5
15
25
4
5
6
7
8
cm
Sepal.Width
Frequency
0
10
20
30
2.0
2.5
3.0
3.5
4.0
cm
Petal.Length
Frequency
0
10
20
30
1
2
3
4
5
6
7
cm
Petal.Width
Frequency
0
5
15
25
35
0.0
0.5
1.0
1.5
2.0
2.5
cm



Correlation Matrix:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1	-0.118	0.872	0.818
Sepal.Width	-0.118	1	-0.428	-0.366
Petal.Length	0.872	-0.428	1	0.963
Petal.Width	0.818	-0.366	0.963	1

3. Principal Component Analysis (PCA)

Method. PCA on standardized numeric features (R prcomp).
Variance explained.

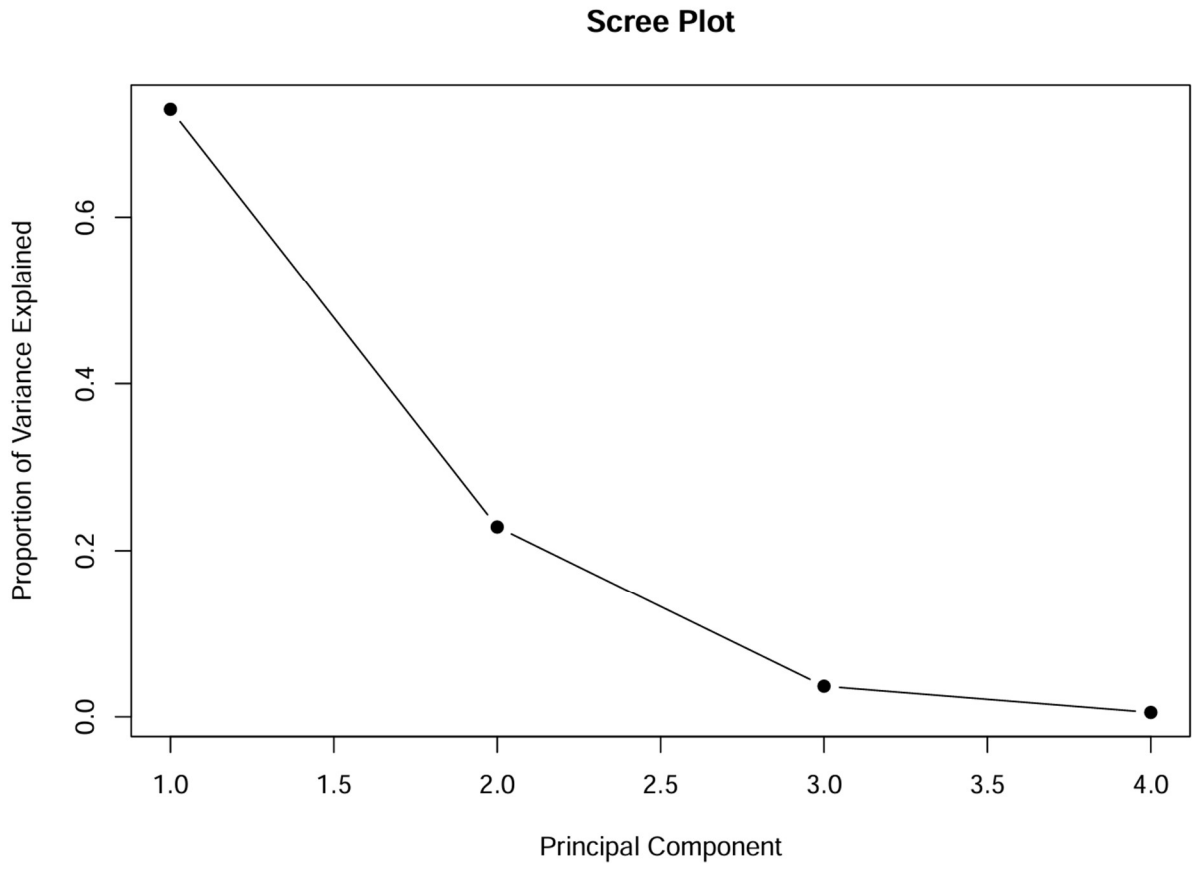
- PC1 = 0.7296, PC2 = 0.2285 → Cumulative = 0.9581 (~95.8%)

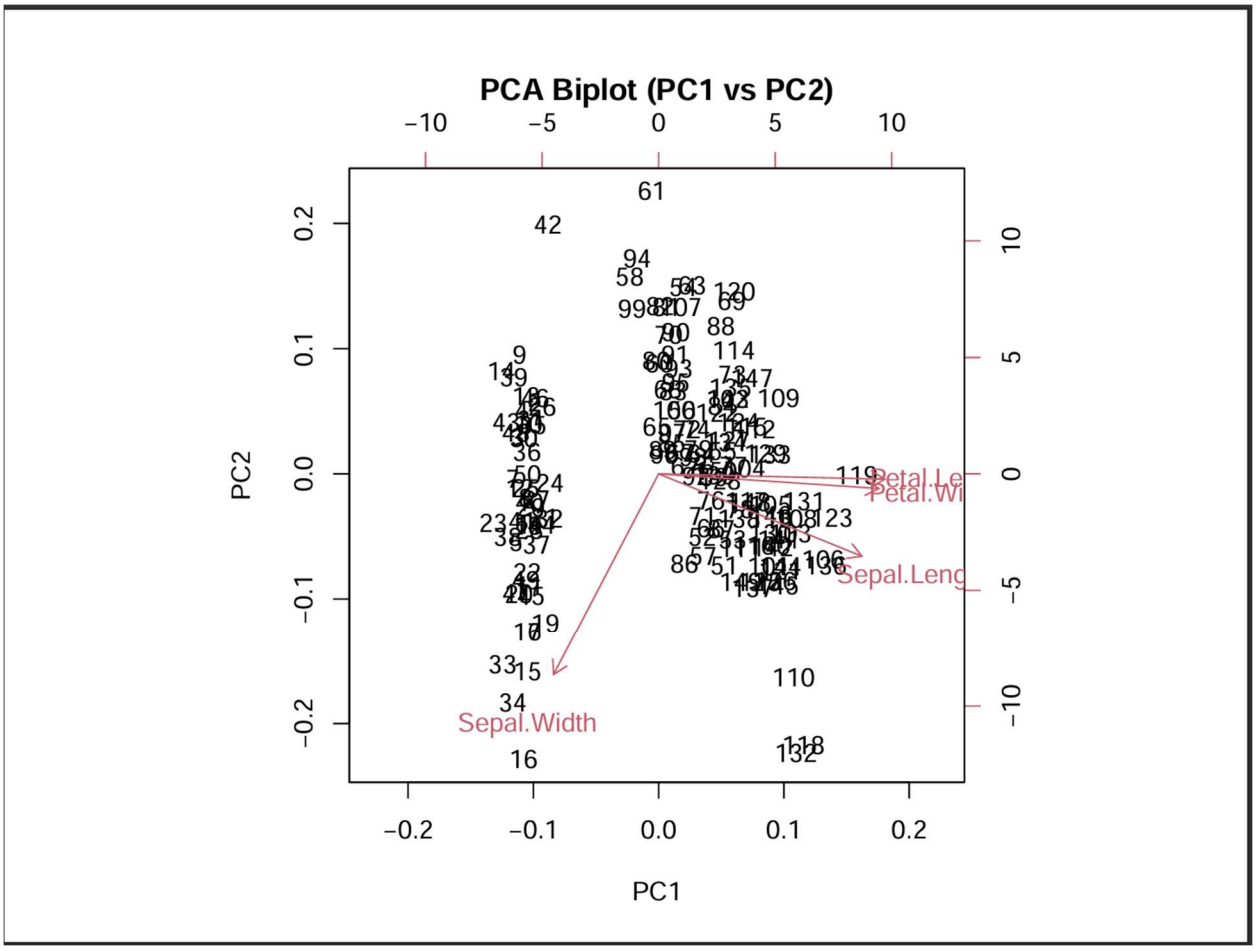
- PC3 = 0.0367, PC4 = 0.0052

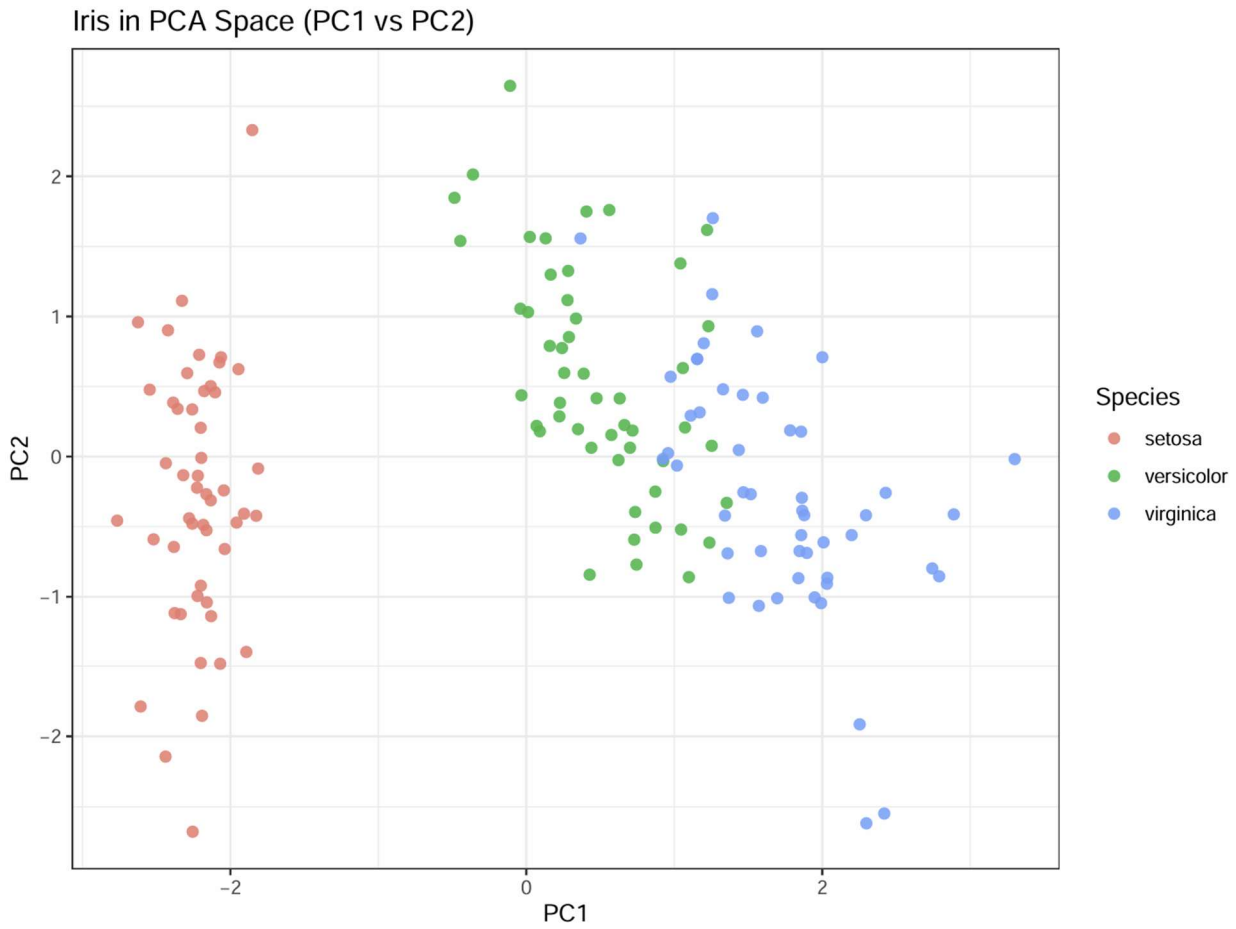
Decision on dimensionality. We retain 2 principal components: the scree plot shows an elbow at PC2, and PC1+PC2 explain 95.8% of total variance.

Interpretation.

- PC1 loads strongly on petal size (Petal.Length/Width) and cleanly separates *setosa* from the other species.
- PC2 contrasts sepal dimensions (notably Sepal.Width vs Sepal.Length).




PCA Biplot (PC1 vs PC2)
PC1
PC2
Sepal.Width
Petal.Le
Petal.Wi
Sepal.Leng



PC	PVE	Cumulative
PC1	0.729624	0.729624
PC2	0.228508	0.958132
PC3	0.036689	0.994821
PC4	0.005179	1

4. Cluster Analysis

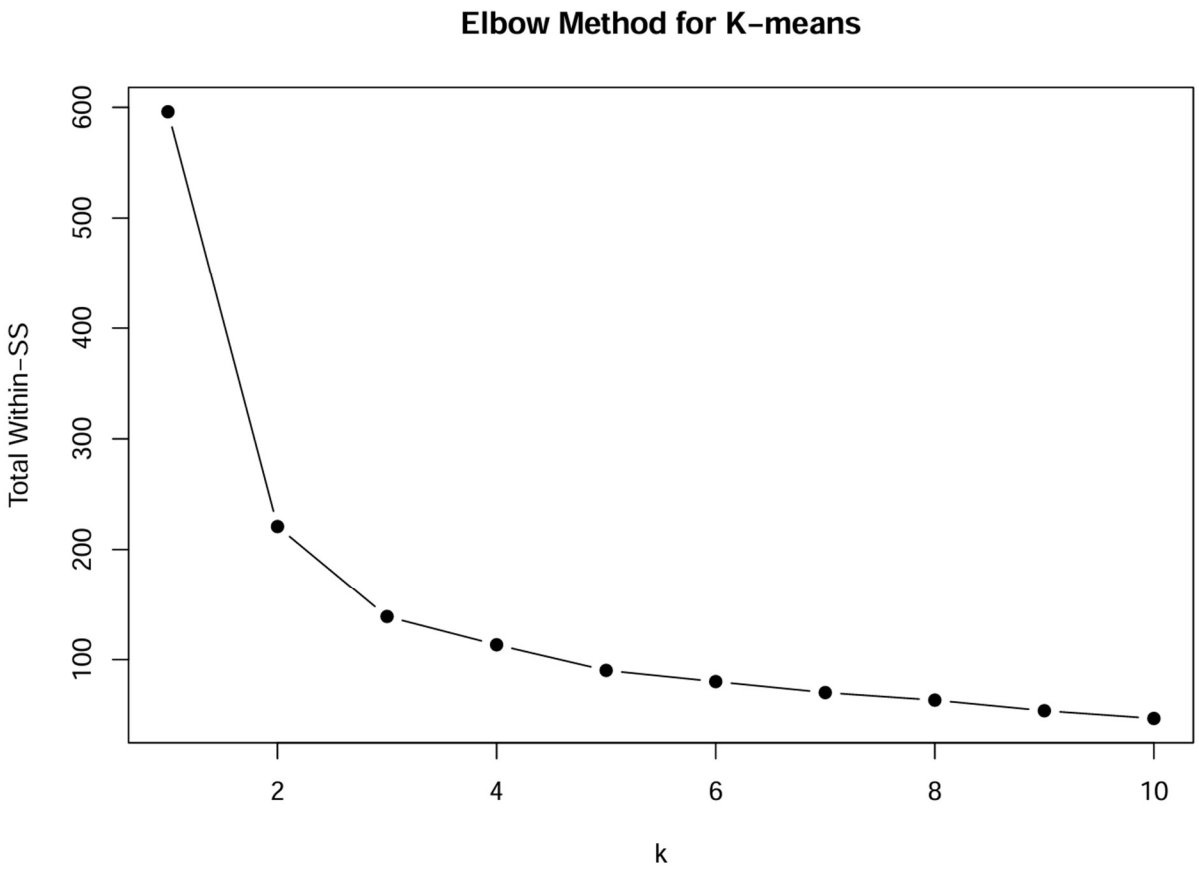
4.1 K-means

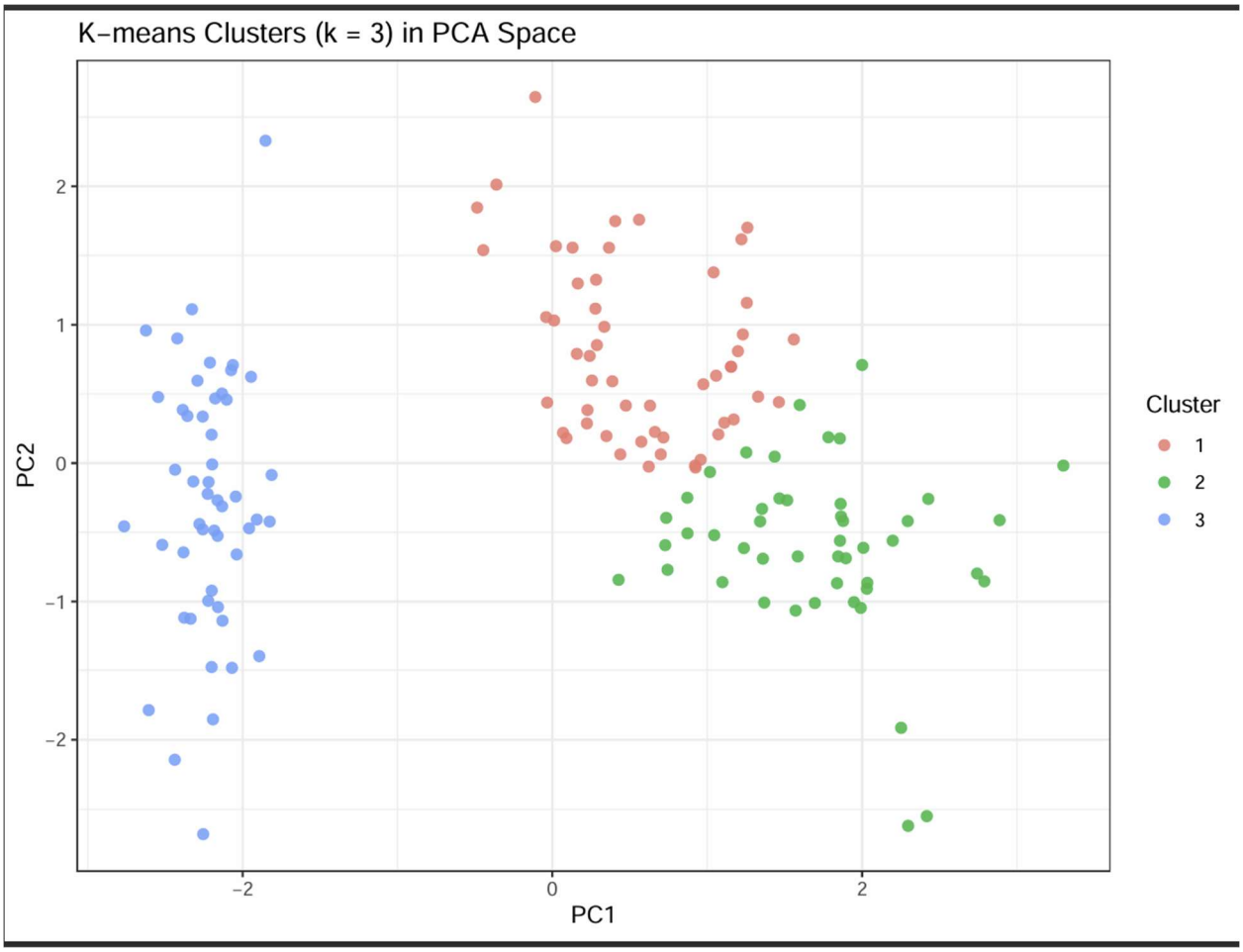
Using the elbow method, k = 3 was selected. Confusion (rows = clusters, cols = species):

- C1: setosa 0, versicolor 39, virginica 14
- C2: setosa 0, versicolor 11, virginica 36
- C3: setosa 50, versicolor 0, virginica 0

Mapping (C3→setosa, C1→versicolor, C2→virginica) gives 125/150 correct (~83.3%).

	setosa	versicolor	virginica
1	0	39	14
2	0	11	36
3	50	0	0



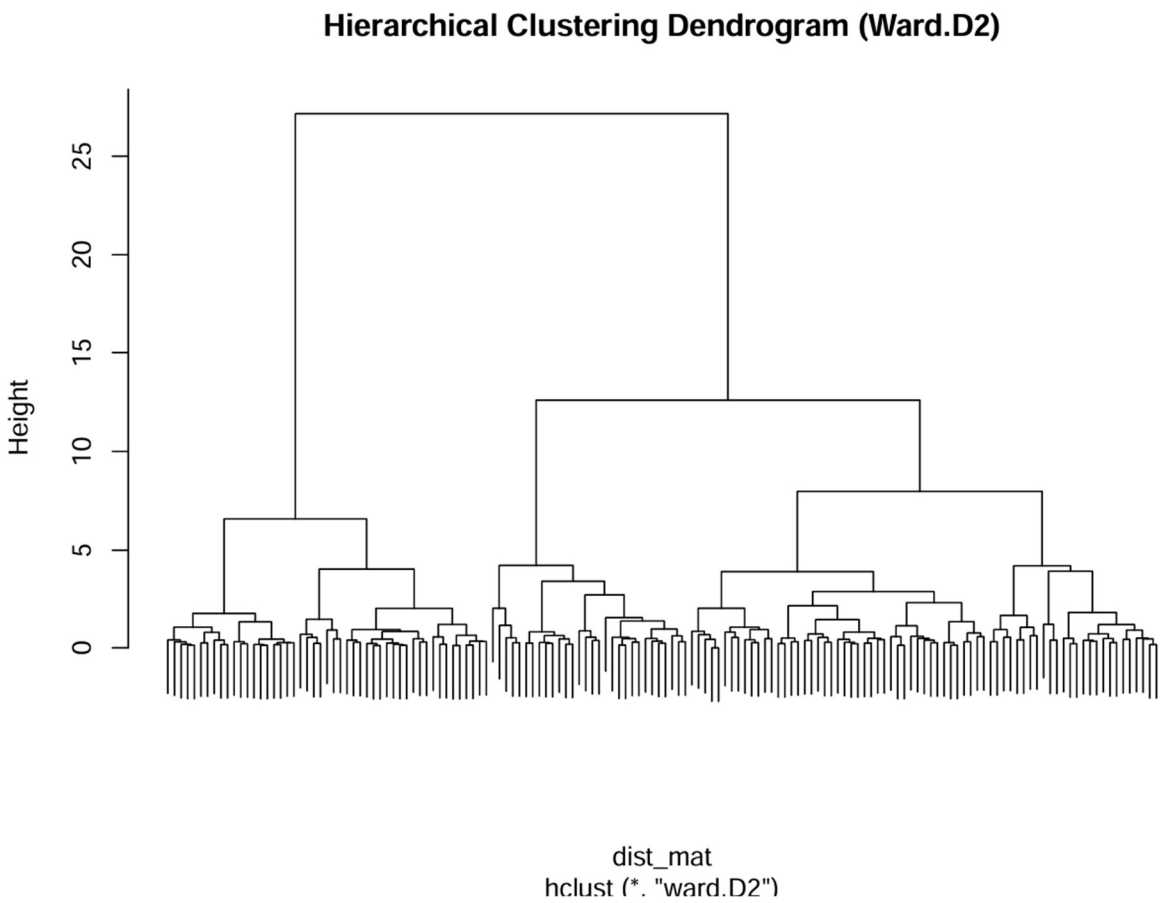


4.2 Hierarchical Clustering (Ward.D2)

Cutting the dendrogram at k = 3 yields:

- Cluster 1: setosa 49, others 0
- Cluster 2: setosa 1, versicolor 27, virginica 2
- Cluster 3: setosa 0, versicolor 23, virginica 48

Assigning (1→setosa, 2→versicolor, 3→virginica) gives 124/150 correct (~82.7%).
Takeaway: *Setosa* is distinctly separable; *versicolor* and *virginica* partially overlap.



	setosa	versicolor	virginica
1	49	0	0
2	1	27	2
3	0	23	48

5. Multivariate Regression

Model (R formula).
Sepal.Length ~ Sepal.Width + Petal.Length + Petal.Width + Species

Key results.

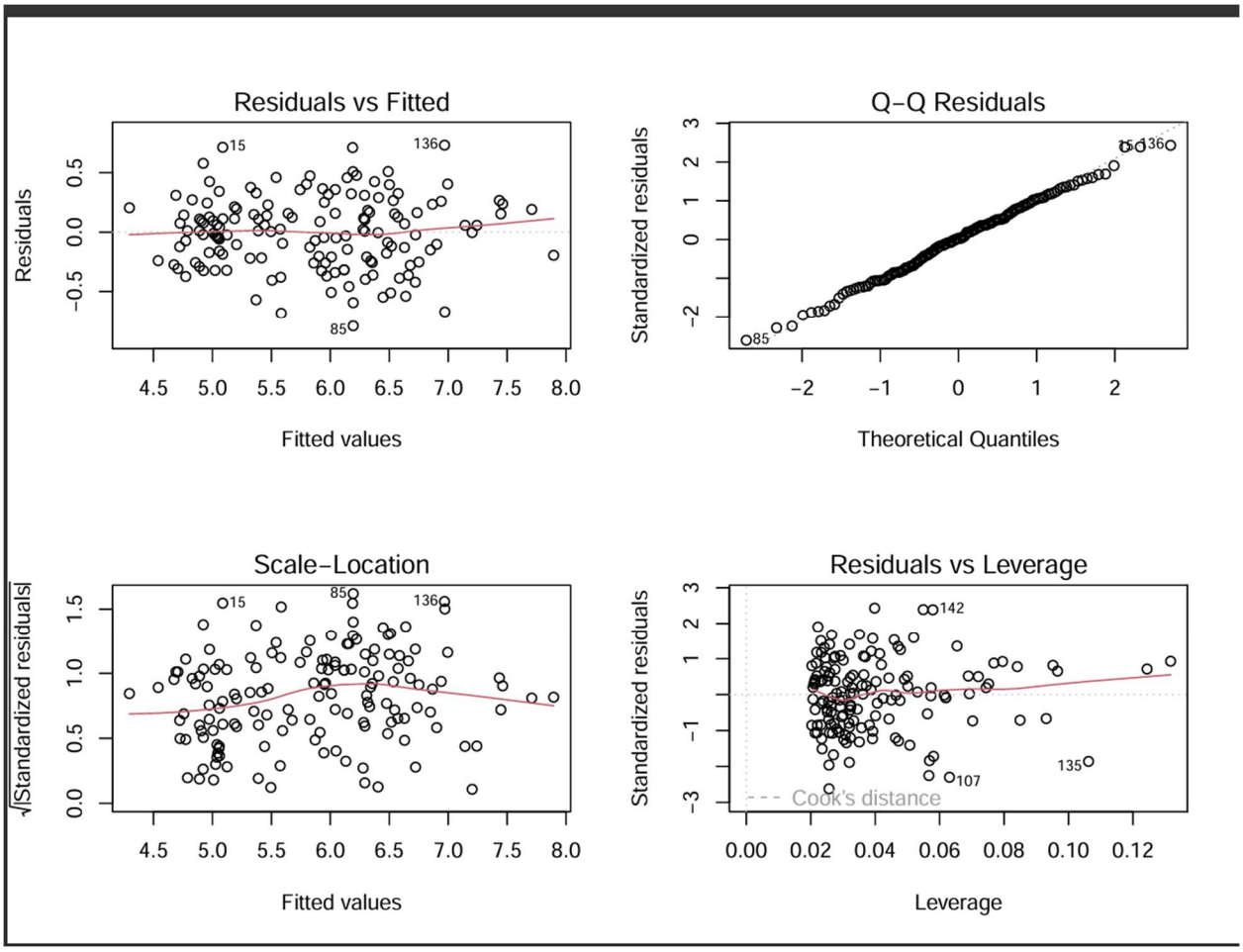
- $R^2 = 0.867$ (Adj. 0.863); $F(5,144) = 188.3$, $p < 2.2e\text{-}16$
- Coefficients (approx):
    - Sepal.Width = +0.496 ($p < 1e\text{-}7$)
    - Petal.Length = +0.829 ($p < 2e\text{-}16$)

    - Petal.Width = −0.315 (p = 0.039)
    - Species indicators vs *setosa*: both significant (p ≈ 0.003)

Multicollinearity (car::vif; GVIF^(1/(2·Df))).

- Sepal.Width 1.49
- Petal.Length 4.81
- Petal.Width 4.58
- Species 2.52
  Petal predictors are near the 5 threshold, indicating notable collinearity—consistent with EDA.

Diagnostics. Residual plots suggest no major violations (approx. normality; variance roughly constant).



Call:

lm(formula = Sepal.Length ~ Sepal.Width + Petal.Length + Petal.Width +

Species, data = df)

Residuals:

Min	1Q	Median	3Q	Max
-0.79424	-0.21874	0.00899	0.20255	0.73103

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)

(Intercept) 2.17127 0.27979 7.760 1.43e-12 ***

Sepal.Width 0.49589 0.08607 5.761 4.87e-08 ***

Petal.Length 0.82924 0.06853 12.101 < 2e-16 ***

Petal.Width -0.31516 0.15120 -2.084 0.03889 *

Speciesversicolor -0.72356 0.24017 -3.013 0.00306 **

Speciesvirginica -1.02350 0.33373 -3.067 0.00258 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3068 on 144 degrees of freedom

Multiple R-squared: 0.8673, Adjusted R-squared: 0.8627

F-statistic: 188.3 on 5 and 144 DF, p-value: < 2.2e-16

	GVIF	Df	GVIF^(1/(2*Df))
Sepal.Width	2.227466	1	1.49247
Petal.Length	23.16165	1	4.812655
Petal.Width	21.0214	1	4.58491
Species	40.03918	2	2.515482

Interpretation: Petal.Length is the strongest positive predictor of Sepal.Length. Including both petal predictors introduces redundancy; a leaner model (e.g., drop Petal.Width) or using PC scores (PC1, PC2) can improve stability with similar fit.

6. Conclusion & Interdisciplinary Note:

- PC1–PC2 capture ~96% of variance; morphology is driven by petal size plus a sepal contrast.
- Unsupervised learning recovers species structure: setosa is separable; versicolor/virginica partially overlap.
- Regression confirms Petal.Length as the strongest positive predictor of Sepal.Length; VIFs indicate redundancy among petal variables.

- Interdisciplinary relevance. The PCA → clustering → regression workflow generalizes to manufacturing quality control (latent dimensions of part measurements), healthcare phenotyping (grouping patients by composite traits), and UX research (segmenting users by behavioral metrics). Ethical practice includes transparent feature selection, reporting uncertainty, and avoiding over-interpreting clusters as discrete “types” when overlap exists